Modelling and Measuring Cosmological Structure Growth

by

Cullan Howlett

This thesis is submitted in partial fulfilment of
the requirements for the award of the degree of
Doctor of Philosophy of the University of Portsmouth.

March 21, 2016

Abstract

Robust measurements of the large scale structure of the universe allow for precise characterisation of its low redshift behaviour and its late time accelerating expansion rate. In particular, Baryon Acoustic Oscillations (BAO) provide a standard ruler with which to measure the expansion rate, whilst Redshift Space Distortions (RSD) allow for tests of General Relativity on cosmological scales. In recent years many surveys have used these probes to investigate the nature of dark energy across a wide range of redshifts with increasing accuracy, culminating in a recent 1% measurement of the BAO scale by Anderson et al. (2014b). Current measurements point towards a consensus cosmological model where dark energy is described only by a Cosmological Constant. However, much of the parameter space available for dark energy models remains unexplored, a point that future surveys such as Euclid (Laureijs et al., 2011), DESI (Levi et al., 2013), LSST (Ivezic et al., 2008) and SKA (Maartens et al., 2015) will attempt to rectify.

This thesis presents work that further confirms the consensus cosmological model using a set of new BAO and RSD measurements at low redshift, whilst also providing tools and techniques to aid in the analysis of next generation datasets.

To begin with, a new code for fast dark matter simulation is presented that can be used to generate large ensembles of accurate mock galaxy catalogues for use in estimating the statistical and systematic errors inherent within large scale structure measurements. The accuracy and speed of this code are tested, where it is found that the new code can reproduce the real-space 2- and 3-point dark matter clustering from a full non-linear N-Body simulation to within 2% and 5% on all scales of interest to BAO and RSD measurements. However each simulation can be run 3 orders of magnitude faster than the corresponding non-linear N-Body run. Several new features are also implemented that will be of use in constructing mock galaxy catalogues for next generation surveys. This code, the algorithms involved and its testing are published in Howlett et al. (2015b)

New measurements of the BAO and RSD signals in a low redshift galaxy sample drawn from the Sloan Digital Sky Survey Data Release 7 are also presented, along with their subsequent cosmological constraints. The simulation code above is first used to generate a set of mock galaxy catalogues based on the low redshift sample, before the

sample and simulations are analysed using the most up-to-date BAO and RSD analysis methods. The procedure for generating the mock catalogues is tested and the clustering of the simulations is found to match that of the data extremely well, even down to scales of $5\,h^{-1}\,\mathrm{Mpc}$. Using the mock catalogues, the BAO and RSD fitting methods are checked for robustness before being used on the data set to obtain a new set of constraints on the expansion rate, equation of state of dark energy and growth rate of structure. In particular, the new BAO measurement completes the low redshift BAO distance ladder and improves current BAO and CMB constraints on the equation of state of dark energy by $\sim 15\%$, to $w_0 = -1.010 \pm 0.081$. This work is published in Ross et al. (2015) and Howlett et al. (2015a).

Finally, a new optimal method for estimating the covariance matrix of the two point clustering of matter is presented, based on a combination of analytic and simulation approaches. This new method can reproduce the covariance matrix estimated from the mock galaxy catalogues simulations used in the rest of this work very well on small scales, in a regime where theoretical estimates of the covariance matrix are extremely difficult to obtain accurately. The benefit of this method is that only simulations that are a fraction of the volume of the full mock galaxy catalogues are required, which in turn means fewer particles are needed to reach the same mass resolution and more simulations (and hence a more precise estimate of the covariance matrix) can be obtained for the same computational cost. The combination of this work and the new fast simulation code presents a much more practical and cost effective way of estimating the covariance matrix.

Table of Contents

List of Tables

List of Figures

List of Acronyms

2dF	2-Degree Field survey
2dFGRS	2-Degree Field Galaxy Redshift Survey
2LPT	2^{nd}-order Lagrangian Perturbation Theory
6dF	6-Degree Field survey
6dFGRS	6-Degree Field Galaxy Redshift Survey
ACT	Atacama Cosmology Telescope
AP	Alcock-Paczynski
BAO	Baryon Acoustic Oscillations
BBN	Big Bang Nucleosynthesis
BOSS	Baryon Oscillation Spectroscopic Survey
CAMB	Code for Anisotropies in the Microwave Background
CDM	Cold Dark Matter
CDF	Cumulative Distribution Function
CLASS	Cosmic Linear Anisotropy Solving System
CLPT	Convolution Lagrangian Perturbation Theory
CMASS	Constant MASS
CMB	Cosmic Microwave Background
COBE	Cosmic Background Explorer
COLA	COmoving Lagrangian Acceleration
CPU	Central Processing Unit

DES	Dark Energy Survey
DESI	Dark Energy Spectroscopic Instrument
DR7	Data Release 7
DR9	Data Release 9
DR11	Data Release 11
EFE	Einstein Field Equation
EOM	Equation Of Motion
EZmocks	Effective Zel'dovich mocks
FFT	Fast Fourier Transform
FFTW	Fastest Fourier Transform in the West
FKP	Feldman-Kaiser-Peacock
FLRW	Friedmanm-Lemaître-Robertson-Walker
FoF	Friends-of-Friends
GADGET	GAlaxies with Dark matter and Gas intEracT
GR	General Relativity
HOD	Halo Occupation Distribution
IC	Initial Conditions
KS	Kolmogorov-Smirnov
L-PICOLA	Lightcone-enabled Parallel Implementation of COLA
LOS	Line-Of-Sight
LOWZ	LOW redshift
LPT	Lagrangian Perturbation Theory
LRG	Luminous Red Galaxy
LSS	Large Scale Structure
MCMC	Markov Chain Monte Carlo

MGS	Main Galaxy Sample
MPI	Message Passing Interface
NFW	Navarro-Frenk-White
NYU-VAGC	New York University Value-Added Galaxy Catalog
Open-MP	OPEN Multi Processing
P3M	Particle-Particle Particle-Mesh
PATCHY	PerturbAtion Theory Catalog generator of Halo and galaxY distributions
PDF	Probability Distribution Function
PINOCCHIO	PIN-pointing Orbit Crossing of Collapsed HIerarchical Objects
PM	Particle-Mesh
PTHALOS	Perturbation Theory HALOS
RSD	Redshift Space Distortions
SDSS	Sloan Digital Sky Survey
SKA	Square Kilometer Array
SPT	South Pole Telescope
VVDS	VIMOS VLT Deep Survey
VIPERS	VIMOS Public Extragalactic Redshift Survey
WMAP	Wilkinson Microwave Anisotropy Probe
ZA	Zel'dovich Approximation

Declaration

Whilst registered as a candidate for the above degree, I have not been registered for any other research award. The results and conclusions embodied in this thesis are the work of the named candidate and have not been submitted for any other academic award.

The work within this thesis has been published in several papers. The L-PICOLA code presented in Chapter 2 is published in Howlett et al. (2015b) with Cullan Howlett primarily responsible for the code development, algorithmic and scientific testing and optimisation, the publication itself, and continued upkeep of the code.

Chapters 3 and 4 are published in two complementary papers (Howlett et al., 2015a; Ross et al., 2015) with Cullan Howlett responsible for the production and testing of the mock catalogues used therein. Select parts of the mock 'pipeline' have also been used by Cullan Howlett for the production of mock catalogues for other datasets (Manera et al., 2015). Additionally, Cullan Howlett was directly responsible for the implementation of the RSD model used, subsequent testing of this on the mock catalogues and for the application of this model to data to obtain measurements of the growth rate before combining this with other datasets as a consistency test of General Relativity.

Finally, Chapter 5 is currently unpublished and is 'work-in-progress'.

Word count: 63,998 words.

Acknowledgements

Of the 60,000 words and 300 pages have gone into this thesis, this is likely the hardest part to write. If I were to thank everyone who should be thanked, this section would likely be longer than the rest of this thesis combined. In lieu of this a few select cases will have to suffice.

First and foremost, on the academic front is of course Will Percival. Without his guidance I would have undoubtedly found myself floundering in an ocean of bad science, surrounded by the flotsam of an academic career. I'd be lying if I said it was all plain sailing, but Will did his best to steer me right and teach me how to do research (I'll drop the nautical puns now, they just sounded good in my head).

Additionally, throughout my three years here, the ICG has had a wealth of talent that I was fortunate enough to tap into. Marc Manera, Ashley Ross, Lado Samushia, Angela Burden and Bridget Falck are just some of the researchers I have had the pleasure of meeting and who have, in some way, helped me conduct the research in this thesis. All of these people have moved on to (hopefully) greener pastures over the years, but their contribution to the research contained in these pages has remained.

At the risk of understating, carrying out a PhD and writing a thesis is difficult. Thankfully there have been many people willing to lend their moral support over the years. To all my fellow PhD students (and other associates), words cannot do justice to the thanks you deserve for helping me though this, always ready to lend an ear to my troubles and share a pint to drown my sorrows. In no particular order, extra special thanks to Claire Le Cras, Harry Wilcox, James Etherington, David Wilkinson, Tim Higgs, Jon Emery, Emma Beynon, Matthew Hull, Daniel Goddard, Xan Morice-Atkinson, Matthew Withers, James Bullock, Adam Baxter and Jimmy Tarr. To those people I have missed out and who deserve to be on this list, thank you too.

Lastly, I would be remiss if I failed to acknowledge the people who deserve the most thanks, my family. Thank you Mum and Dad for encouraging me from a young age to pursue what I want and for giving me the best possible opportunity for doing that. Thank you Connor for keeping me grounded and being my oldest and truest friend. You're the real reason I visit home. And thank you Grandad and Auntie Sharon for your continual

support and for reminding me that if all else fails I still have a home to go to. The last couple of years have been rocky for all of us, but I hope the future has some gentler times in store.

And to Nana, whose generosity and teachings will continue to shape me forever, and whose love will remain with me for the rest of my days, if somehow you are reading this, this thesis is for you.

Chapter 1

Introduction

Over the last century, a cohesive picture of the nature of our Universe has emerged, in the form of the Hot Big Bang model. This model consists of two main tenets. The first of these is that the geometry and dynamics of the universe are mathematically described by solutions to General Relativity (GR; Einstein 1916) derived using the Friedman-Robertson-Lemaître-Walker metric (FLRW; Friedmann 1924; Lemaître 1927; Robertson 1935; Walker 1935). Within this framework, the geometry of the universe is sourced by the energy and momentum contained therein. The second idea behind the Hot Big Bang model, is that the energy-density of the universe is described by the ΛCDM model, which consists of four components:

Radiation - Any relativistic matter, including photons and Standard Model neutrinos.

Baryonic Matter - Any non-relativistic matter from within the Standard Model of Particle Physics.

Dark Matter - Extremely weakly interacting matter, from beyond the Standard Model of Particle Physics. This is only observable via its gravitational influence.

Λ - The cosmological constant, responsible for the accelerated expansion of the universe.

Along with the assumption of the Cosmological Principle, which is satisfied by using the FLRW metric, the combination of GR and ΛCDM form the concordance model of Cosmology. The Cosmological Principle posits that the universe is both statistically homogenous and statistically isotropic. This is an extension of the Copernican Principle, that we do not exist in some *special* location within the universe. Whilst this assumption is certainly incorrect on small scales there is evidence to support *statistical* homogeneity and isotropy on scales $\gtrsim 300\,h^{-1}\,\mathrm{Mpc}$ within the observable universe (Maartens et al., 1995; Clarkson & Maartens, 2010; Scrimgeour et al., 2012) including the incredible uniformity of the Cosmic Microwave Background radiation (CMB; Penzias & Wilson 1965; Smoot et al. 1992; Mather et al. 1994). However, the presence of large scale anomalies

in both the CMB and the large scale distribution of galaxies in the universe continues to challenge this assumption (Cruz et al., 2007; Clowes et al., 2013; Horváth et al., 2014; Planck Collaboration et al., 2014c) and the commonly accepted theory of inflation, which will be touched on later, requires that the universe may be inhomogeneous on scales well outside the observable universe.

Regardless of the accuracy of the assumption of the Cosmological Principle, there exists overwhelming observational evidence in support of the Hot Big Bang model, including:

- The expansion of the universe, which is currently believed to be undergoing a period of acceleration (Slipher, 1912; Hubble, 1929; Riess et al., 1998; Perlmutter et al., 1999).
- The existence of the CMB, which can be explained by the decoupling of radiation and baryonic matter some time after the Hot Big Bang, at which time expansion has caused the coupled plasma of the early universe to cool (Penzias & Wilson, 1965; Smoot et al., 1992; Bennett et al., 2013; Planck Collaboration et al., 2014a).
- The existence of elements heavier than hydrogen in the very early universe, which in the Hot Big Bang were created via Big-Bang Nucleosynthesis (BBN; Alpher et al. 1948; Coc et al. 2004).
- The observed clustering of baryonic matter, forming galaxies and large scale structures (LSS). After decoupling the baryonic matter can fall into the gravitational potential wells created by the dark matter and form galaxies. The origin of the initial density perturbations from which these form is a result of the inflationary paradigm.

This chapter provides an in depth introduction to the current, 'concordance' model of cosmology, so called due to the abundance of observational evidence pointing towards a single cosmological model, defined by only a few parameters. This concordance model fits all current observations. To start with, in Section 1.1, an introduction to General Relativity and the FLRW metric will be given. The solutions to GR under this metric are presented and used to explain the evolution of the universe since the Hot Big Bang.

The combination of GR and the FLRW metric provides a general picture of the evolution of background geometry of the universe, but these alone cannot explain the presence of LSS in the Universe. Section 1.2 presents a detailed picture of the how structure in the universe evolves over cosmic time from the initial seeds left by quantum fluctuations prior to Inflation. This section will detail the evolution of the density perturbations arising from Inflation up to the present day, traversing the radiation, matter and Λ dominated regimes that exist in the standard model of cosmology.

From thereon the rest of the chapter will be dedicated to the observational study of the universe and the evidence that led to the Hot Big Bang model being named the 'concordance' cosmlogical model. Within the framework of the ΛCDM model, GR and the FLRW metric, there exist several key, 'free' parameters, such as the relative abundance of matter and the equation of state of dark energy, which determine the exact cosmology and evolution of the universe. Section 1.3 will give a brief overview of the different cosmological probes used to measure these parameters, the current constraints from these probes, and the questions about the concordance model that still remain in the light of these measurements.

In the final section, 1.4, additional detail will be given on the two main probes of the late time large scale structure of the universe, Baryon Acoustic Oscillations (BAO) and Redshift Space Distortions (RSD). The origin of these phenomena will also be covered in this chapter. Overall this section will contain a detailed foundation on which much of the work presented in this thesis resides and will be referred to frequently in later chapters.

In this chapter the Einstein Summation Convention and natural units, where the speed of light $c = 1$, are adopted. Many general physics textbooks and cosmology resources have been used to prepare this introduction, in particular the textbooks of Dodelson (2003); Liddle & Lyth (2000); Liddle (2003); Lyth & Liddle (2009); Mukhanov (2005); Peacock (1999); Peebles (1980); Weinberg (2008). The lecture notes found at `http://cosmologist.info/teaching/` have also been extremely helpful.

1.1 The Concordance Model of Cosmology

1.1.1 General Relativity

In the early part of the 20th Century, one of the predominant theories of modern physics, Einstein's theory of General Relativity (Einstein, 1916) was established. Within this framework, gravity is defined as the curvature of a four-dimensional spacetime. Matter distorts the geometry of the spacetime, which in turn sources the gravitational force felt between two objects.

This theory is based on a set of underlying principles:

- **The Equivalence Principle:**
  The inertial mass of an object is equal to its gravitational mass, such that all objects small enough not to contribute to the gravitational field itself have the same acceleration under gravity. As a direct consequence of this, an observer in an infinitesimally small (such that there are no *tidal* forces) local inertial frame freely falling under gravity observes the same physical laws as those in an inertial frame without a gravitational field.

- **The General Principle of Relativity:**
  All observers are equivalent in terms of the physical laws they observe. This is another extension of the Copernican Principle.

- **The Principle of General Covariance:**
  This extension of the General Principle of Relativity states that the laws of physics should be tensor equations, such that they are invariant under a coordinate transformation.

- **The Correspondence Principle:**
  Any new physical law must recover the well-known older physical laws in the appropriate limits. For example, in the absence of gravity, General Relativity should reduce to Special Relativity, and in the presence of a weak gravitational field, General Relativity must reduce to Newtonian Gravity.

The Geometry of the Universe

On the back of these fundamental principles, Einstein provided a mathematical description of how the curvature of spacetime is encoded by its matter content and how the curvature of spacetime causes matter to gravitate. The infinitestimal line-element connecting two four-vectors within a general spacetime is given by

$$ds^2 = g_{\mu\nu} dx^\mu dx^\nu \tag{1.1}$$

where $g_{\mu\nu}$ is the spacetime metric. In a generically curved spacetime, differentiation can no longer be performed in the usual way as the differential is no longer invariant under coordinate transformations. Satisfying the Principle of General Covariance requires the introduction of a covariant derivative which *is* invariant under coordinate transformations. The covariant derivative $a_{\mu;\sigma}$ of a covariant four-vector, a_μ with respect to a contravariant four-vector x^σ is given by

$$a_{\mu;\sigma} \equiv a_{\mu,\sigma} - \Gamma^\nu_{\mu\sigma} a_\nu \tag{1.2}$$

where $a_{\mu,\sigma} = \partial a_\mu / \partial x^\sigma$ is the ordinary derivative, and $\Gamma^\nu_{\mu\sigma}$ are the Christoffel symbols.

The Christoffel symbols are an affine connection, which allows one to transport vectors on a spacetime while keeping them parallel to the connection itself. They can be defined using the metric as

$$\Gamma^\nu_{\mu\sigma} = \Gamma^\nu_{\sigma\mu} \equiv \frac{1}{2} g^{\nu\rho} \left(g_{\rho\mu,\sigma} + g_{\rho\sigma,\mu} - g_{\mu\sigma,\rho} \right). \tag{1.3}$$

If the spacetime is flat, $g_{\mu\nu} = \eta_{\mu\nu} = \mathrm{diag}(1, -1, -1, -1)$ and $\Gamma^\nu_{\mu\sigma} = 0$.

Because of the curvature of spacetime, covariant derivatives do not commute. This property can be used to define the curvature, via an object called the Riemann tensor

$$a^{\mu}{}_{;\alpha;\beta} - a^{\mu}{}_{;\beta;\alpha} = -R^{\mu}{}_{\nu\alpha\beta}a^{\nu}. \tag{1.4}$$

Using the definition of the Christoffel symbol, this can be rewritten

$$R^{\mu}{}_{\nu\alpha\beta} \equiv \Gamma^{\mu}_{\beta\nu,\alpha} - \Gamma^{\mu}_{\alpha\nu,\beta} + \Gamma^{\mu}_{\alpha\rho}\Gamma^{\rho}_{\beta\nu} + \Gamma^{\mu}_{\beta\rho}\Gamma^{\rho}_{\alpha\nu}. \tag{1.5}$$

In the presence of spacetime curvature, transport of vectors around a closed loop does not return the vectors to their initial orientation and as such bringing two vectors together may result in a different answer depending on the order of operation, i.e., which coordinate the vectors are brought together in first. The Riemann tensor quantifies this difference. In a flat spacetime there is no difference and so $R^{\mu}{}_{\nu\alpha\beta} = 0$, which is apparent as the Christoffel symbols are all zero.

Finally the Einstein tensor can be defined via suitable contractions of the Riemann tensor to the Ricci tensor $R_{\mu\nu} \equiv R^{\alpha}{}_{\mu\alpha\nu}$ and the Ricci scalar $R \equiv R^{\mu}{}_{\mu} \equiv g^{\mu\nu}R_{\mu\nu}$,

$$G_{\mu\nu} \equiv R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R \tag{1.6}$$

This is the tensor used in General Relativity to quantify the curvature of the universe. Although it is not the only option, it is the simplest possible combination of the Riemann tenor and its contractions that satisfies the principles set out previously. Why the Riemann tensor or Ricci scalar cannot be used alone will become apparent once the matter content of the universe is defined mathematically.

The Matter Content of the universe

In General Relativity the curvature of spacetime is related to the matter therein. As such the tensor describing the geometry must be related to a tensor describing the energy density. The energy-momentum tensor, $T^{\mu}{}_{\nu}$ does just this. It contains all the information about the energy and momentum of material within the universe.

The Cosmological Principle demands that $T^{\mu}{}_{\nu}$ be a perfect fluid to conserve homogeneity and isotropy. Under these conditions the energy-momentum tensor can be generically written as

$$T^{\mu}{}_{\nu} = (\rho + P)U^{\mu}U_{\nu} - P\delta^{\mu}_{\nu} \tag{1.7}$$

where ρ is the mass density of the fluid and P is its pressure. U^{μ} is the relative four-velocity between the fluid and the reference frame of the observer. For an observer comoving with the fluid $U^{\mu} = \mathrm{diag}(1, 0, 0, 0)$ and $T^{\mu}{}_{\nu} = \mathrm{diag}(\rho, -P, -P, -P)$.

Energy-momentum conservation can be written neatly as

$$T^{\mu}{}_{\nu;\mu} = 0 \tag{1.8}$$

and in an empty space without vacuum energy $T^{\mu}{}_{\nu} = 0$.

The Einstein Field Equations

Now that mathematical expressions for the curvature of the universe and its matter content have been defined, all that's left is to combine these to form the Einstein Field Equations (EFEs)

$$G_{\mu\nu} = \frac{8\pi G}{c^4} T_{\mu\nu} \tag{1.9}$$

The constant on the right-hand side of the equation is derived using the Principle of Correspondence, which demands that the solutions to these equations reduce to Newtonian gravity in the presence of weak gravitational fields. The properties of the energy-momentum tensor enforce that the left-hand side of the EFEs cannot be only the Ricci tensor as this is not covariantly conserved ($R^{\mu\nu}{}_{;\nu} \neq 0$). The Riemann tensor alone cannot be used as not only is this a rank-4 tensor and incompatible with the energy-momentum tensor, but in a spacetime containing no matter or energy, the EFEs would imply that the Riemann tensor is also 0 and the spacetime would be flat. However gravity from nearby objects does exist in empty space.

1.1.2 The Friedmann-Lemaître-Robertson-Walker metric

As a set of 10 coupled, non-linear differential equations, the EFEs are extraordinarily difficult to solve for a general matter distribution without some underlying assumption about the form of the spacetime metric. Under the assumption of the Cosmological principle, that the universe is homogenous and isotropic, the most general metric one can formulate is the FLRW metric (Friedmann, 1924; Lemaître, 1927; Robertson, 1935; Walker, 1935),

$$ds^2 = dt^2 + a^2(t)[dr^2 - S_k^2(r)(d\theta^2 + sin^2\theta d\phi^2)], \tag{1.10}$$

where

$$S_k(r) = \begin{cases} \sqrt{k^{-1}}\sin(r\sqrt{k}) & k > 0 \\ r & k = 0 \\ \sqrt{|k|^{-1}}\sinh(r\sqrt{|k|}) & k < 0 \end{cases}. \tag{1.11}$$

$a(t)$ is the scale factor of the universe, which measures its size as a function of time. It is common practice to normalise this such that $a = 1$ is the present-day scale factor. The function $S_k(r)$ defines the intrinsic curvature of the universe. There are three potential geometries that satisfy the conditions of isotropy and homogeneity: flat, spherical and hyperbolic, which correspond to $k = 0$, $k > 0$ and $k < 0$ respectively. These are often called flat, closed and open universes respectively.

Adopting the FLRW metric allows the EFEs to be solved analytically for a perfect isotropic fluid. These solutions were first presented by Friedmann (1922).

1.1.3 FLRW Solutions to the EFEs

The following pair of equations for the evolution of the geometry and matter content of the universe over cosmic time can be derived from the combination of the FLRW metric, energy-momentum tensor and EFEs

Friedmann Equation

Adopting the FLRW metric, such that $g_{\mu\nu} = \mathrm{diag}(1, -a^2(t), -a^2(t), -a^2(t))$ and solving the 13 non-zero Cristoffel symbols allows the Ricci tensor to be written as

$$R^t{}_t = 3\frac{\ddot{a}}{a} \quad ; \quad R^r{}_r = R^\theta{}_\theta = R^\phi{}_\phi = \frac{\ddot{a}}{a} + \frac{2\dot{a}^2 + 2kc^2}{a^2}. \tag{1.12}$$

A 'dot' ($\dot{}$) denotes the derivative with respect to time. The Ricci scalar is

$$R = 6\left(\frac{\ddot{a}}{a} + \frac{\dot{a}^2 + k}{a^2}\right). \tag{1.13}$$

Finally, plugging these into the time-time EFE, one arrives at the Friedman equation

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G\rho}{3} - \frac{k}{a^2}. \tag{1.14}$$

This details how the intrinsic curvature and the density of the components within the universe affects its expansion rate.

Continuity Equation

The continuity equation can be most easily derived using the conservation of the energy-momentum tensor. Taking the $\nu = 0$ component

$$T^\mu{}_{0;\mu} = T^\mu{}_{0,\mu} + \Gamma^\mu_{\alpha\mu}T^\alpha{}_0 - \Gamma^\alpha_{0\mu}T^\mu{}_\alpha = 0. \tag{1.15}$$

The only non-zero component of the tensor $T^\alpha{}_0$ is when $\alpha = 0$, and under the FLRW metric the only relevant non-zero Christoffel symbols are $\Gamma^\alpha_{0\alpha} = \dot{a}/a$. Substituting in these Christoffel symbols, one arrives at the continuity equation,

$$\dot{\rho} + 3\frac{\dot{a}}{a}(\rho + P) = 0, \tag{1.16}$$

which details how density is conserved in an expanding universe.

1.1.4 The Time Evolution of the Universe

The matter content of the universe, consisting of four distinct contributions, has been shown to source the gravitational potential by affecting the geometry of spacetime itself.

Using the Friedmann and fluid equations enables one to study the evolution of the components of the universe over cosmic time. The four contributors can be treated as a perfect fluid, with an equation of state $p = -\omega\rho c^2$. The value of ω differs for different types of fluid, $\omega = 0$, $-1/3$ and 1 for non-relativistic matter (both baryonic and dark), radiation and the cosmological constant respectively.

Assuming that any one of these components dominates over the others and substituting their equation of state into the fluid equation results in

$$\rho \propto \begin{cases} a^{-3} & \mathrm{Non-relativistic\,matter} \\ a^{-4} & \mathrm{Radiation} \\ constant & \mathrm{Cosmological\,Constant} \end{cases} \tag{1.17}$$

for the different components of the energy-momentum tensor. As would be expected, the density of non-relativistic matter falls off in proportion to the volume of the universe. Radiation on the other hand falls off more sharply. This is because the density also decreases due to radiation losing energy as the universe expands, which introduces an additional factor of 'a'. Finally the density of the cosmological constant does not change as a function of scale factor at all, and remains constant (the actual value is $\rho_\Lambda = \Lambda c^2/8\pi G$).

A more general solution to the Friedmann equation is that the universe contains a mixture of these components. A convenient way to rescale the Friedmann equation to highlight this is to define the critical density

$$\rho_c(a) = \frac{3H^2(a)}{8\pi G}, \tag{1.18}$$

which is the density required to ensure a flat universe at some scale factor. The definition of the Hubble parameter $H(a) = \dot{a}/a$, which has a present day value H_0, has been used here. This will be explained in more depth in the next section. The density of each of the components can be normalised by the critical density, such that $\Omega_i = \rho_i/\rho_c$. Writing the density parameters as a function of their present day values and substituting these into the Friedmann equation allows one to rewrite this as

$$H^2(a) = H_0^2 E^2(a) = H_0^2 \left(\frac{\Omega_{m,0}}{a^3} + \frac{\Omega_{r,0}}{a^4} + \frac{\Omega_{k,0}}{a^2} + \Omega_{\Lambda,0} \right), \tag{1.19}$$

where $\Omega_{m,0}$, $\Omega_{r,0}$ and $\Omega_{\Lambda,0}$ are the normalised present day densities for non-relativistic matter, radiation and the cosmological constant respectively. $\Omega_{k,0} = kc^2/H_0 = 1 - \Omega_{m,0} - \Omega_{r,0} - \Omega_{\Lambda,0}$ is the curvature density, the energy density due to the intrinsic curvature of the universe, which due to the normalisation used must be equal to the remainder when all other contributions to the density are subtracted from 1.

For a general mixture of components to the energy-momentum tensor, Eq. 1.19 provides a simple way to track the time evolution of these universe, as long as the present

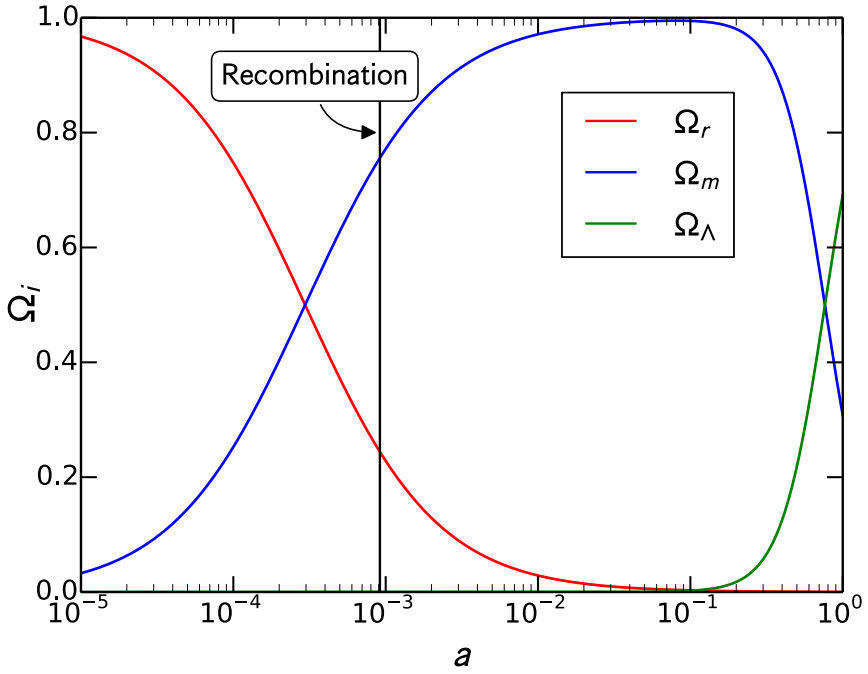


Figure 1.1: The evolution of the radiation, matter and cosmological constant density as a function of scale factor. The universe starts off radiation dominated up until matter-radiation equality (the cross-over between the red and blue lines) after which matter starts to dominate. At late-times the universe then enters a phase of Λ domination. The solid black line indicates the epoch of recombination, where the photons decouple from the baryons and free-stream away.

day values are known (or the values at some known time in the past). Many studies have performed measurements of these values using a variety of probes. The current consensus values from the combination of CMB, BAO and Supernovae data are given in Section 1.3. Figure 1.1 shows the evolution of the density parameters as a function of scale factor based on these measured present day values.

1.1.5 Universal Expansion and Redshift

The Friedmann equation also explains the observation first made by Hubble (1929) that the recession velocity of galaxies is a function of their distance from an observer, leading to the theory of an expanding universe. The recession velocity of an object in direction $\boldsymbol{r}$ is given by

$$\boldsymbol{v} = \frac{d\boldsymbol{r}}{dt} = \frac{|\dot{\boldsymbol{r}}|}{|\boldsymbol{r}|}\boldsymbol{r} = \frac{\dot{a}}{a}\boldsymbol{r}. \tag{1.20}$$

Relating this to Hubble's law $\boldsymbol{v} = H\boldsymbol{r}$, shows the origin of the Hubble parameter defined previously. The value of the Hubble parameter today is often given by $H_0 =$

$100h\,\mathrm{kms}^{-1}\mathrm{Mpc}^{-1}$. A variety of probes of the expansion rate of the universe (Conley et al., 2011; Riess et al., 2011; Freedman et al., 2012; Tammann & Reindl, 2013; Aubourg et al., 2014; Efstathiou, 2014) have enabled tight constraints to be put on the exact value of h (H_0). Current state of the art constraints combining CMB, BAO and Supernovae data from the Planck Collaboration et al. (2015a) give the consensus value $H_0 = 67.74 \pm 0.46\,\mathrm{kms}^{-1}\mathrm{Mpc}^{-1}$.

One remarkable consequence of the expansion of the universe is that massless particles lose energy as they travel though it. Indeed, it is this property that enabled the expansion of the universe to be recognised. This phenomenon can be seen by using the key property that massless photons travel on null geodesics, i.e., they obey $ds^2 = 0$. Plugging this into the FLRW metric and using isotropy such that $d\theta = d\phi = 0$,

$$\frac{dt}{a} = dr. \tag{1.21}$$

If one wanted to find the total time taken for a light ray to travel some distance from $r = 0$ to $r = D_c$, they could integrate this expression between the time of emission and the time of receiving (t_e and t_r respectively). Here a trick can be used, which is that the object that emitted the light ray is stationary in *comoving* coordinates (it is receeding only because of the scale factor $a(t)$). As such light emitted and received a short time later ($t_e + \Delta t_e$ and $t_r + \Delta t_r$) will travel the *same* comoving distance. That is,

$$\int_{t_e}^{t_r} \frac{dt}{a} = \int_{t_e+\Delta t_e}^{t_r+\Delta t_r} \frac{dt}{a}. \tag{1.22}$$

Rearranging the limits of this integral and taking $\Delta t_{e,r}$ to be small

$$\frac{dt_e}{a(t_e)} = \frac{dt_r}{a(t_r)}. \tag{1.23}$$

Finally, if, instead of two light rays, the above expression is taken for successive peaks of a single light wave, where the time between the two peaks is proportional to the wavelength λ

$$\frac{\lambda_r}{\lambda_e} = \frac{a(t_r)}{a(t_e)} = 1 + z. \tag{1.24}$$

In the later part of the above expression, the redshift, z, of the light has been defined. If the light is received by an observer at the present day $a(t_r) = 1$ the scale factor at which the light was emitted is related to its redshift by

$$a(t_e) = \frac{1}{1+z}. \tag{1.25}$$

The final remark in this section is on the relation between redshift and comoving distance. It is common for the redshift of an object in the universe, which is the observable quantity, to be used as a proxy for the distance of that object from an observer. Strictly

speaking this conversion is only possible if the Hubble parameter is known exactly as a function of time (or redshift/scale factor). Performing the integral of Eq. 1.21

$$D_c = \int_{t_e}^{t_r} \frac{cdt}{a} = c\int_{a(t_e)}^{1} \frac{da}{aH(a)} = c\int_0^z \frac{dz}{H(z)}, \tag{1.26}$$

where D_c is the comoving distance of the object and the speed of light has been reintroduced to recover the correct units. The last two expressions are after a change of basis has been performed by making use of the relation between the time derivative of the scale factor, the Hubble parameter and the redshift. This latter expression is useful for converting the measured redshift of an object to a comoving distance, as it is relatively easy to write the Hubble parameter as a function of the energy-density of the universe, from the four main components listed previously.

Several other distance measures can also be defined based on the comoving distance:

- **Transverse Comoving Distance:**
  The distance between two objects at constant redshift z, seperated by some angle θ is given by θD_m, where D_m is the transverse comoving distance. This is related to the comoving distance and curvature via

$$D_m = \begin{cases} \frac{H_0}{c}\sqrt{\Omega_k^{-1}}\sinh\left(D_c\frac{H_0}{c}\sqrt{\Omega_k}\right) & \Omega_k > 0 \\ D_c & k = 0 \\ \frac{H_0}{c}\sqrt{|\Omega_k|^{-1}}\sin\left(D_c\frac{H_0}{c}\sqrt{|\Omega_k|}\right) & \Omega_k < 0 \end{cases} \tag{1.27}$$

- **Angular Diameter Distance:**
  The angular diameter distance, D_A, is the ratio of an objects true size to the angle it subtends on the sky (in radians). The angular diameter distance is related to the transverse comoving distance,

$$D_A = \frac{D_m}{1+z}. \tag{1.28}$$

  This distance scale is commonly used in the discussion of *standard rulers* and will be revisted later in this chapter during the discussion of measuring BAO in the LSS, whose known physical size can be used to measure the cosmology of the universe.

- **Luminosity Distance:**
  Finally, the luminosity distance, D_L, can be used to relate an object's intrinsic luminosity, L, to the flux from that object measured by an observer, S. The ratio between the two gives the luminosity distance

$$D_L = \sqrt{\frac{L}{4\pi S}}. \tag{1.29}$$

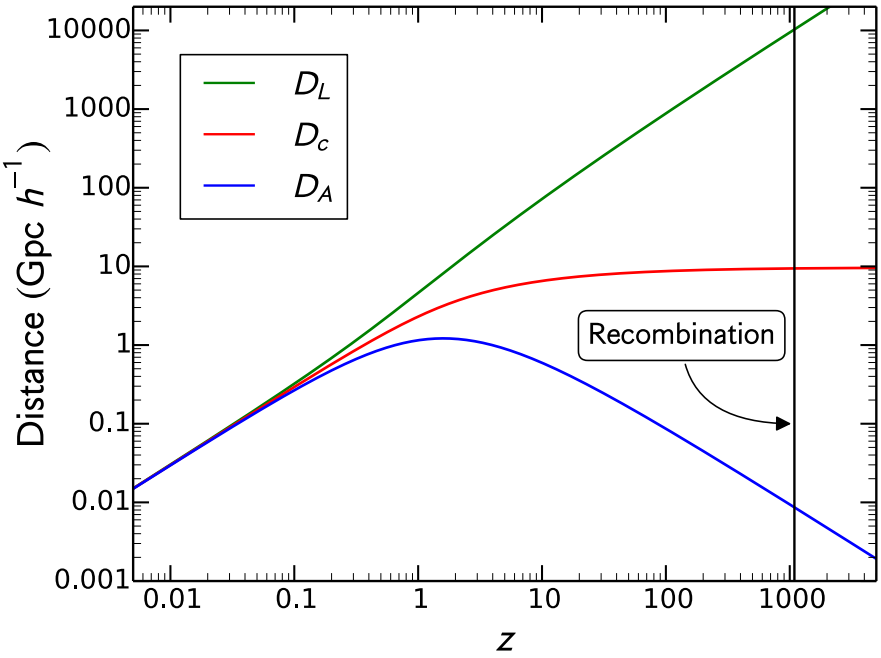


Figure 1.2: The comoving, angular diameter and luminosity distance as a function of redshift for the consensus cosmology given in Eq 1.84. The almost perfectly flat cosmology of the universe means that the transverse comoving distance would be indistinguishable from the comoving distance, $D_m = D_c$, and so is not included in the plot. As is commonly done, the distances are quoted with the Hubble constant factored out, making them independent of the exact value of H_0.

This is often used in the context of *standard candles* whose known intrinsic luminosity can be used to constrain the cosmology of the universe. The luminosity distance is related to the angular diameter and transverse comoving distances through

$$D_L = (1+z)D_M = (1+z)^2 D_A. \tag{1.30}$$

Figure 1.2 shows these different distance measures as a function of the redshift from the observer.

1.2 Evolution of Perturbations

The previous section has shown how the geometry and components of the universe are intimately linked via the EFEs and how they evolve over time. The key assumption in this section was that of the Cosmological Principle, that the universe is both homogeneous and isotropic. However the presence of structure in the universe, from life itself to clusters of galaxies, seems at odds with this assumption. It is only on very large scales that this holds

true. The solution to this apparent conflict is that the background geometry of the universe *is* homogeneous and isotropic, and the structure we see at the present day has evolved from tiny quantum fluctuations that seeded the otherwise uniform background prior to an epoch of extremely rapid expansion at the emergence of the universe. This section will present the linear evolution of these perturbations in the background geometry of the universe from their inception to the present day.

From hereon, all equations are written in terms of the conformal time, $d\eta = dt/a$ and conformal Hubble parameter $\mathcal{H} = Ha$. In the interest of pedagogy, flatness is also assumed, however as shown above this is a reasonable assumption.

1.2.1 Inflation and the Primordial Perturbations

The smooth homogeneous and isotropic background that arises from the solutions to GR under the FLRW metric cannot explain the presence of large scale structure in the universe. Furthermore, the discovery of the extremely uniform CMB in the 1960's led scientists to question how parts of the universe that are so disparate today, that could never have been in causal contact under the Hot Big Bang model, have a temperature uniformity that could only arise from thermal equilibrium. This is called the *Horizon Problem*. A solution to this is that very, very early on in the history of the universe, there was a period of rapid, accelerated expansion, called Inflation (Starobinskiĭ, 1979; Kazanas, 1980; Einhorn & Sato, 1981; Guth, 1981; Albrecht & Steinhardt, 1982; Linde, 1982). This period of Inflation enlarged causally connected patches of the universe to well beyond the current Hubble radius.

Although originally hypothesised to explain the Horizon Problem, Inflation also solves several other outstanding problems with the Hot Big Bang model. The first of these is the *Flatness Problem*. Current measurements give a universe that is very close to flat ($\Omega_k \approx 0$). As shown in 1.1.4, the curvature density of the universe evolves in proportion to a^{-2}. In comparison to the a^{-3} and a^{-4} dependence of the matter and radiation, this means that for the curvature to not dominate today the curvature density at the beginning of the universe must have been exceptionally small, or exactly zero. This presents a serious case of 'fine-tuning'. Inflation provides a natural solution to this as the rapid, accelerated expansion of the early universe forces the intrinsic curvature of the observable universe to be very small.

Another problem is the *Relic Problem*. Phase transitions in the early universe, where the universe cools below some temperature threshold for unification of the fundamental forces, give rise to very massive particles. The mass of these particles is on the order of the energy scale at which the phase transition occurs and so they would contribute massively to the matter density. However these particles are not detected in the present day matter

distribution. One solution to this is that there is no such unification in the early universe and so no Relic particles. Inflation also presents a solution to this by causing such a rapid expansion in the universe as to dilute the contribution to the matter density from relic particles.

The final problem that can be explained by Inflation is the presence of density fluctuations in the homogeneous background. Small-scale quantum fluctuations prior to Inflation get greatly enlarged and become the seeds of the structure we see today.

There still exists many unsolved questions about the period of accelerated expansion in the early universe, including the nature of the 'Inflaton' field driving the expansion. Regardless of this, the ability of the theory of Inflation to provide solutions to Horizon, Flatness and Relic problems, as well as presenting an intuitive origin for primordial density perturbations, has seen it widely adopted into the Hot Big Bang model.

1.2.2 Linear Perturbations in the Newtonian Gauge

A fully robust treatment of the evolution of perturbations over cosmic time requires solving versions of the EFEs where both the metric and energy-momentum tensor are perturbed. Furthermore, as the components of the energy-momentum tensor are not actually perfect fluids, their behaviour depends on more than just their density and pressure. As such, solving the perturbed EFEs requires solving the relativistic Boltzmann equation for the fluid. This differential equation describes the time evolution of the full phase space distribution (composed of 3 spatial coordinates and 3 momenta) of a given fluid. Solving the perturbed EFEs alongside the Boltzmann equations for each contributor to the energy-momentum tensor can only be done numerically and is computationally demanding.

In the context of this work, and in the interest of providing only a pedagogical, analytic overview of the evolution of perturbations up to the present day, it is sufficient instead to use the assumption of Newtonian perturbations. By the Correspondence Principle solutions obtained under this assumption should be applicable for all perturbations within the Hubble radius and for weakly gravitating systems.

Assuming only scalar perturbations, in the Newtonian gauge the perturbed metric can be written as

$$ds^2 = a(\eta)^2[(1+2\Psi)d\eta^2 - (1-2\Phi)\delta_{ij}dx^i dx^j)]. \tag{1.31}$$

Ψ is the usual Newtonian gravitational potential, whilst Φ is the perturbation to the spatial curvature itself.

For a perfect fluid, the perturbation to the stress-energy tensor, $\tilde{T}^{\mu}_{\ \nu} = T^{\mu}_{\ \nu} + \delta T^{\mu}_{\ \nu}$ can be decomposed as

$$\tilde{T}^{0}_{\ 0} = \rho + \Delta\rho \tag{1.32}$$

$$\tilde{T}^i_{\ 0} = (P+\rho)v^i \tag{1.33}$$

$$\tilde{T}^i_{\ j} = -(P+\Delta P)\delta^i_j + \Pi^i_{\ j}, \tag{1.34}$$

where i, j indicate the three spatial indices, $\Delta\rho$ and ΔP are the density and pressure perturbations, v^i is the velocity in the direction of the i^{th} spatial coordinate and $\Pi^i_{\ j}$ is the anisotropic stress tensor. For matter the anisotropic stress tensor is zero and will be neglected from now on. This assumption also enforces $\Psi = \Phi$.

Perturbed Conservation and Euler Equations

Conservation of energy allows for the following expressions, $\tilde{T}^{0\mu}_{\ \ ;\mu} = 0$ and $\tilde{T}^{i\mu}_{\ \ ;\mu} = 0$. Solving these and extracting the linear expressions for the perturbations results in

$$\Delta\rho' + 3\mathcal{H}(\Delta\rho + \Delta p) + (\rho + P)\nabla\cdot\boldsymbol{v} + 3\Phi'(\rho + P) = 0 \tag{1.35}$$

and

$$\boldsymbol{v}' + \mathcal{H}\boldsymbol{v} + \frac{P'}{\rho + P}\boldsymbol{v} + \frac{\nabla\Delta p}{\rho + P} + \nabla\Phi = 0 \tag{1.36}$$

where $'$ denotes a derivative with respect to conformal time, $\eta = \int_0^t dt/a(t)$. These are the Continuity and Euler equations. They can be rewritten by using the Fluid equation and defining the overdensity $\delta = \Delta\rho/\rho$ and sound speed $c_s^2 = \Delta P/\Delta\rho$, such that

$$\delta' + 3\mathcal{H}\delta\left(c_s^2 - \frac{P}{\rho}\right) + \left(1 + \frac{P}{\rho}\right)\left(\nabla\cdot\boldsymbol{v} + 3\Phi'\right) = 0 \tag{1.37}$$

and

$$\boldsymbol{v}' + (1 - 3c_s^2)\mathcal{H}\boldsymbol{v} + \frac{c_s^2\nabla\delta}{1 + \frac{P}{\rho}} + \nabla\Phi = 0 \tag{1.38}$$

Perturbed Poisson Equation

A third equation can be derived by solving the perturbed EFEs, under the perturbed metric and energy-momentum tensor. Upon calculating the required perturbed Christoffel symbols, Riemann and Ricci tensors and Ricci scalar to linear order, the $0-0$ component of the Einstein tensor can be shown to be

$$G_{00} = 3\mathcal{H}^2 + 2\nabla^2\Phi - 6\mathcal{H}\Phi'. \tag{1.39}$$

Combining this with the $0-0$ component of the energy-momentum tensor, solving the corresponding EFE and making use of the background solution, the Friedmann equation, results in

$$\nabla^2\Phi = 4\pi G a^2\rho\delta + 3\mathcal{H}(\Phi' + \mathcal{H}\Phi). \tag{1.40}$$

Next, making use of the spatial solutions to the EFEs allows the solution

$$\Phi' + \mathcal{H}\Phi = -4\pi G a^2(\rho + P)\boldsymbol{v}. \tag{1.41}$$

Finally, substituting this into the solution from the $0-0$ EFE

$$\nabla^2\Phi = 4\pi G a^2 \rho\delta + 3\mathcal{H}(\rho + P)\boldsymbol{v} = 4\pi G a^2 \rho\bar{\delta}. \tag{1.42}$$

The last expression follows from recognising that $\bar{\delta} = \delta + 3\mathcal{H}(1+\frac{P}{\rho})\boldsymbol{v}$ is just the comoving overdensity. The familiar form of this final equation belies its nature. This is simply the linear order perturbed Poisson equation.

On subhorizon scales, the combination of the perturbed Continuity, Euler and Poisson equations allow for analytic solutions to the linear growth of matter perturbations in the background expanding universe. These will be used to derive the time evolution of radiation and matter at different epochs during the universe. This evolution can be split into several distinct epochs, which are shown in Fig. 1.1: the pre-Recombination era, where the photons and baryons are tightly coupled such that their evolution is intimately connected; post-Recombination, where the universe has expanded and cooled sufficiently to allow the photons to decouple from the baryons and free-stream away; and the late-time dark energy dominated universe. These will be tackled in individual sections below. Throughout, subscripts γ, b, c and m will denote contributions from photons, baryons, cold dark matter and all matter (baryons plus cold dark matter) respectively.

1.2.3 Perturbations Pre-Recombination

Perturbations in the Photon-Baryon Fluid

Prior to recombination, the photons and baryons are tightly coupled due to Compton scattering. The high temperatures of the early universe mean that the protons and electrons are unbound from one another. Photons scattering off this charged plasma results in a tight coupling between the photons, baryons and electrons, such that the velocity of the two is virtually equivalent.

Conservation of momentum requires that that the momentum of the photon-baryon fluid is given by the sum of the momenta of the photons and non-relativistic baryons, such that the total momentum can be expressed

$$(\rho + P)\boldsymbol{v} = (\rho_\gamma + P_\gamma)\boldsymbol{v}_\gamma + (\rho_b + P_b)\boldsymbol{v}_b = \frac{4}{3}(1+R)\rho_\gamma \boldsymbol{v}_\gamma. \tag{1.43}$$

The last equality stems from making use of the equations of state for the photons and baryons, assuming tight coupling between the two, and making the convenient definition

$$R \equiv \frac{3\rho_b}{4\rho_\gamma}, \tag{1.44}$$

which encapsulates the density ratio between the photons and baryons.

The expression for the momentum conservation of the photon-baryon fluid allows the Euler equation for this fluid to be expressed purely in terms of the photon properties only

(or baryon properties, as will be shown subsequently), with a suitable rescaling based on R

$$\boldsymbol{v}'_\gamma + \frac{R}{1+R}\mathcal{H}\boldsymbol{v}_\gamma + \frac{1}{1+R}\frac{\nabla\delta_\gamma}{4} + \nabla\Phi = 0. \tag{1.45}$$

Again, making use of the photon equation of state, the continuity equation for the photons has the form

$$\delta'_\gamma + \frac{4}{3}\nabla\cdot\boldsymbol{v}_\gamma + 4\Phi' = 0. \tag{1.46}$$

Fourier transforming this pair of equations, such that the gradient terms become functions of the wavevector k, differentiating the continuity equation and substituting this into the Euler equation results in the second-order formula for the density of the photons (Peebles & Yu, 1970; Doroshkevich et al., 1978; Hu & Sugiyama, 1995; Hu & White, 1996)

$$\delta''_\gamma + \frac{R}{1+R}\mathcal{H}\delta'_\gamma + \frac{1}{3(1+R)}k^2\delta_\gamma = -4\Phi'' - \frac{4R}{1+R}\mathcal{H}\Phi' - \frac{4k^2}{3}\Phi. \tag{1.47}$$

A similar equation can be written in terms of the baryon properties (Eisenstein et al., 2007a),

$$\delta''_b + \frac{R}{1+R}\mathcal{H}\delta'_b + \frac{1}{3(1+R)}k^2\delta_b = -3\Phi'' - \frac{3R}{1+R}\mathcal{H}\Phi' - k^2\Phi. \tag{1.48}$$

In both cases the forms of these equations are recognisable. They are the equations for a forced, damped harmonic oscillator with sound speed

$$c_s^2 = \frac{1}{3(1+R)}. \tag{1.49}$$

In the absence of baryons, sound waves would travel though the plasma of the early universe at the standard rate of $c_s^2 = 1/3$. However, the presence of baryons makes the fluid heavier and slows down the resultant sound waves. Physically the equations governing the evolution of the density of fluid describe how, on subhorizon scales, the gravitational potential (which 'forces' the oscillations) causes the photon-baryon fluid to compress. This is turn raises the temperature and pressure of the medium and the resultant outwards pressure force causes rarefaction. These *acoustic oscillations* continue until the epoch of recombination, when the decoupling of photons and baryons means that they no longer act as a single fluid. However the sound waves present prior to recombination leave an imprint in the photons and baryons *after* recombination. This will be discussed later.

Perturbations in the Cold Dark Matter

Unlike the baryons, which tightly couple to the photons prior to recombination, the cold dark matter component of the universe interacts with the other components only via the gravitational potential. As such, for non-interacting, non-relativistic dark matter $P_c = 0$

and $c_s^2 = 0$. Additionally, the rapid oscillations of the photon-baryon fluid described previously mean that these actually contribute very little to the growth of the gravitational potential. During radiation domination, $\mathcal{H}$ decays with conformal time and the rapid oscillations in the photon density mean that the density is, on average, reasonably constant. As such the radiation contribution to the gravitational potential is vanishing due to the expansion of the universe.

With this in mind, the Continuity, Euler and Poisson equation for the Cold Dark Matter can be written, respectively, as

$$\delta_c' + \nabla \cdot \boldsymbol{v}_c = -3\Phi', \tag{1.50}$$

$$\boldsymbol{v}_c' + \mathcal{H}\boldsymbol{v}_c = -\nabla\Phi, \tag{1.51}$$

$$\frac{3}{2}\Omega_c(\eta)\mathcal{H}^2\delta_c = \nabla^2\Phi, \tag{1.52}$$

where the relation $4\pi G\rho_c a^2 = \frac{3}{2}\Omega_c(\eta)\mathcal{H}^2$ has been used.

Combining these three equations into a single, second-order differential equation as previously and neglecting the time derivatives of the Newtonian potential, which are negligible on subhorizon scales, gives

$$\delta_c'' + \mathcal{H}\delta_c' = \frac{3}{2}\Omega_c(\eta)\mathcal{H}^2\delta_c. \tag{1.53}$$

During radiation domination $\Omega_c(\eta) \approx 0$ and the CDM overdensity grows logarithmically at most. Matter domination starts very early on even in the pre-recombination era, and as such a solution is required that works both before and after the era of matter-radiation equality. To accomplish this the above equation can be recast under the change of variable

$$y = \frac{a}{a_{eq}} \approx \frac{\rho_m}{\rho_\gamma}. \tag{1.54}$$

The resultant formula is the Meszaros equation (Meszaros, 1974)

$$\frac{d^2\delta_c}{dy^2} + \frac{2+3y}{2y(y+1)}\frac{d\delta_c}{dy} - \frac{3}{2y(y+1)}\delta_c = 0. \tag{1.55}$$

This equation has solutions

$$\delta_c = A\left(1+\frac{3}{2}y\right) + B\left[\left(1+\frac{3}{2}y\right)\ln\left[\frac{\sqrt{1+y}+1}{\sqrt{1+y}-1}\right] - 3\sqrt{1+y}\right]. \tag{1.56}$$

In the radiation dominated era, $y << 1$ and

$$\delta_c \approx (A-3B) - B\ln\left(\frac{y}{4}\right), \tag{1.57}$$

i.e., the CDM overdensity grows logarithmically as expected. In the matter dominated regime, there are two contributions to the solution, a growing mode proportional to y (or equivalently, the scale factor a) and a decaying mode proportional to $y^{-3/2}$,

$$\delta_c \approx A\left(1+\frac{3}{2}y\right) + \frac{4}{15}\frac{B}{y^{3/2}}. \tag{1.58}$$

Physically this means that during radiation domination the growth of matter perturbations is suppressed by the radiation. The CDM can only grow slowly as the scale factor of the universe is controlled by the radiation and the baryons are stuck oscillating alongside the photons. The fact that structure is observed in the CMB indicates that the matter must come to dominate very early on in the history of the universe, long before recombination. For this to be possible it is hence necessary that the amount of CDM must significantly outweigh the amount of baryonic material.

1.2.4 Perturbations Post-Recombination

Radiation Perturbations

The density fluctuations in the photon-baryon fluid prior to recombination give rise to temperature anisotropies which are present in the Cosmic Microwave Background. The Cosmic Microwave Background is composed of the photons that free-streamed away from the surface of last scattering at recombination, after decoupling themselves from the baryons. At the time of decoupling, whilst the universe had cooled sufficiently for the protons and electrons to combine, the universe was still relatively young and hot. The temperature of the photons was $\sim 3000K$. However in the intervening $\sim 13\mathrm{Gyr}$ since the the epoch of recombination, the expansion of the universe has caused the photons emitted at the surface of last scattering to redshift and lose energy. As such, their temperature as measured today is significantly less, $T = 2.7255 \pm 0.0006K$ (Mather et al., 1994).

The CMB follows an almost perfect black-body spectrum. However, high precision measurements of the CMB temperature across the sky reveal anisotropies of around 1 part in 10^5 (Smoot et al., 1992; Bennett et al., 1994; Fixsen et al., 1997; Bennett et al., 2013; Planck Collaboration et al., 2014a) . These are caused, for the most part, by the acoustic oscillations present in the pre-recombination photon-baryon plasma. The damped, forced oscillations in the density of the photon-baryon fluid give rise to scale dependent temperature fluctuations in the CMB which can then be measured and modelled as a function of their scale.

A temperature anistropy at some position $\hat{\boldsymbol{n}}$ on the sky is given by

$$\Theta(\hat{\boldsymbol{n}}) = \frac{\Delta T(\hat{\boldsymbol{n}})}{T(\hat{\boldsymbol{n}})}. \tag{1.59}$$

As the primordial perturbations that ultimately source the temperature fluctuations are assumed to be drawn from a Gaussian with zero mean, then $\langle \Theta(\hat{\boldsymbol{n}}) \rangle = 0$ and the relevant statistic of interest of is the power spectrum of the fluctuations at some angular scale θ

$$C(\theta) = \langle \Theta(\hat{\boldsymbol{n}}) \Theta(\hat{\boldsymbol{n}'}) \rangle. \tag{1.60}$$

It is common practice to decompose this into spherical harmonics, with amplitude $a_{\ell m}$

$$\Theta(\hat{\boldsymbol{n}}) = \sum_{\ell=1}^{\infty} \sum_{m=-\ell}^{\ell} a_{\ell m} Y_{\ell m}(\hat{\boldsymbol{n}}), \tag{1.61}$$

such that the measured power spectrum of temperature fluctuations can be rewritten

$$C_\ell = \langle a_{\ell m} a^*_{\ell m} \rangle. \tag{1.62}$$

In the above expression there is an implicit sum over the m, arising from the fact that, for a given ℓ, the variance of $a_{\ell m}$ is constant. This in turn means that for a measurement of C_ℓ there are a finite number of modes that can be averaged over, $2\ell + 1$ to be exact. In the limit of low ℓ this means that there is a finite precision that can be reached in measurements of C_ℓ due to the presence of only a few modes stretching across the full-sky. This is the cosmic variance limit.

The anisotropy power spectrum of the CMB has been measured by several surveys over the last 30 years, including measurements by the COBE (Fixsen et al., 1997), SPT (Keisler et al., 2011), ACT (Das et al., 2011), WMAP (Hinshaw et al., 2013) and Planck (Planck Collaboration et al., 2014b) surveys. Figure 1.3 shows the anisotropy power spectrum from a combination of the WMAP, ACT and SPT data (Hinshaw et al., 2013). Modelling of the shape of the power spectrum gives us the strongest constraints on the cosmology of our universe from any single probe.

The shape of the temperature fluctuation power spectrum is determined by several factors. Primarily, the acoustic oscillations resulting from the photon-baryon plasma prior to recombination imprints a series of peaks and troughs on the power spectrum. The density perturbation in the photons at the point of decoupling, i.e., the solution to Eq. 1.47, has the form $\delta_\gamma(k) \propto \cos(k r_s(\eta))$, where $r_s(\eta)$ is the radius of the sound horizon at decouping

$$r_s(\eta) \equiv \int_0^{\eta_{dec}} d\eta' c_s(\eta'). \tag{1.63}$$

The oscillatory nature of the density perturbations sets up oscillations in the temperature anisotropy power spectrum, with wavelength dependent on the radius of the sound horizon at decoupling. This in turn means that the wavelength of the oscillations is dependent on the conformal time at decoupling and the sound speed up until this point. Furthermore, density perturbations of different scales (different k's) enter the horizon and become causally connected at different times. Once a k-mode has entered the horizon it can begin to grow with a rate related to the sound speed. This effect means that successive peaks in the k-space acoustic oscillations have different amplitudes based on the relative abundances of baryons and photons when it entered the horizon.

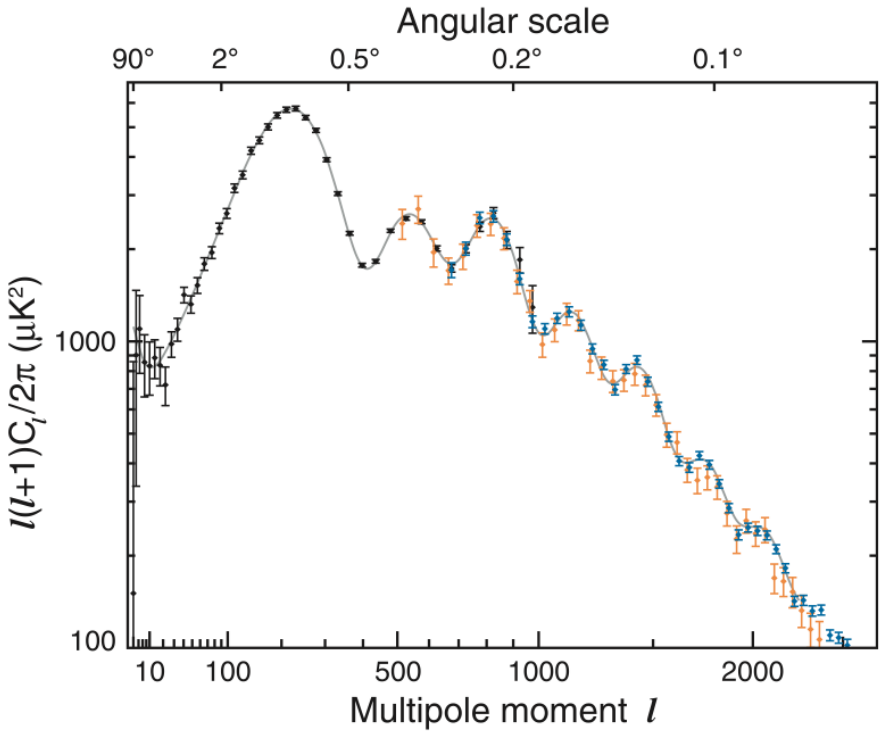


Figure 1.3: The measured CMB temperature anisotropy power spectrum from a combination of WMAP (black; Hinshaw et al. 2013), ACT (orange; Das et al. 2011) and SPT (blue; Keisler et al. 2011) data, along with the best-fit cosmological model to the WMAP data, shown in grey. This plot is taken from (Hinshaw et al., 2013).

There are also other effects that affect the shape of the CMB power spectrum. The acoustic oscillations that are the solutions to the damped, forced harmonic oscillator equation governing the photon-baryon fluid are further damped due to the imperfect coupling between the photon and baryons prior to decoupling. In deriving Eq. 1.47 it was assumed that the baryons and photons are perfectly coupled, the Thomson scattering rate between the two is infinite, such that their velocities are equal. In reality the photons must travel some finite distance before scattering which leads to diffusion, or 'Silk', damping (Silk, 1968) of small scale features in the power spectrum. This same phenomenon also affects the acoustic oscillations seen in the baryons.

Finally, on the largest scales, the derivation of Eq. 1.47 is incomplete as it neglects perturbations in the metric which can dominate for regions outside the horizon. These perturbations are caused by differences in the gravitational potential in non-causally connected patches of the universe sourced by differing matter densities. Photons at the surface of last scattering get redshifted as they climb out of these gravitational wells, introducing anisotropies in the large scale CMB. This is called the Sachs-Wolfe effect (Sachs & Wolfe, 1967). In the non-Newtonian regime these perturbations in the metric are actually the primary source of anisotropies in the CMB and as such this dominates the low-ℓ power spectrum. A secondary form of this, the Integrated Sachs-Wolfe effect (Rees & Sciama, 1968; Crittenden & Turok, 1996), arises due to the CMB photons losing energy as they pass through expanding gravitational potential wells between the observer and the surface of last scattering.

Overall, the combination of acoustic oscillations, Sachs-Wolfe effect and Silk damping means that full shape of the CMB power spectrum is very sensitive to the baryonic, CDM, curvature and cosmological densities and Hubble constant. In particular, $\Omega_b h^2$, $\Omega_m h^2$, Ω_k and Ω_Λ are very well determined by modelling the power spectrum of temperature anisotropies in the CMB. Increasing $\Omega_b h^2$ decreases the sound speed prior to recombination and as such increases the wavelength of the acoustic oscillations and the relative amplitude of the first peak compared to the second peak. Conversely, increasing the CDM density decreases the wavelength of the acoustic oscillations by reducing the radiation density driving the oscillations. The strength of the large scale anisotropies is also increased compared to the small scale anisotropies due to the increased effect of gravitational potential wells outside the horizon. Ω_k directly affects the projection of physical scales to angular scales on the surface of last scattering. In a closed universe a feature of fixed physical size is projected onto larger angular scales and as such the position of the peaks in the CMB power spectrum is shifted to larger scales. This effect is similar to that of the Cosmological Constant, although the degeneracy can be broken by the Integrated Sachs-Wolfe effect which, in the presence of a Cosmological Constant, enhances

the large scale anisotropies and causes the relative amplitude of the peaks to change.

Numerical predictions of the CMB power spectrum rely on solving the full set of Einstein Field and Boltzmann equations for a given cosmological model to recover the expected temperature anisotropy on the sky $\Theta(\hat{\boldsymbol{n}})$, given some initial perturbation. This is then decomposed into spherical harmonics and C_ℓs computed from the multipoles of the temperature fluctuations. This presents a rather complex task but there are several codes available to do this, such as CMBFAST (Seljak & Zaldarriaga, 1996), CAMB (Lewis et al., 2000) and CLASS (Lesgourgues, 2011).

Matter Perturbations

The imprint of the sound waves on the baryons after recombination takes the form of a conventional density contrast. After decoupling the sound horizon of the photon-baryon fluid leaves a characteristic scale at which the baryons will have been compressed, leading to an overdensity of baryonic material. In fact even after decoupling, the baryons still retain some of the momentum they had whilst coupled to the photons, which causes the sound wave to continue to propagate till shortly after the epoch of recombination. The epoch at which the sound horizon of the baryons is defined is called the baryon drag epoch, so labelled as the baryons are *dragged* along for some time after recombination. This in turn means that the sound horizon for the photons as measured from the CMB and for the baryons will be subtly different. Eisenstein & Hu (1998) provide an analytic formula for the sound horizon at the redshift of the baryon-drag epoch

$$r_d = \int_0^{t(z_d)} = c_s(1+z)dt = \frac{2}{3k_{eq}}\sqrt{\frac{6}{R_{eq}}}\ln\frac{\sqrt{1+R_d}+\sqrt{R_{eq}+R_d}}{1+\sqrt{R_{eq}}}, \tag{1.64}$$

where R_{eq}, R_d are the photon-baryon density ratio at the matter-radiation equality and baryon-drag epochs, and

$$k_{eq} = \sqrt{2\Omega_{m,0}H_0 z_{eq}} \tag{1.65}$$

is the scale of the horizon at the redshift of matter-radiation equality, z_{eq}.

Much like the CMB anisotropies, the sound horizon of the *Baryon Acoustic Oscillations* can provide a wealth of information, particularly in the sense that, as this depends only on the relative densities of photons and baryons at the matter-radiation equality/baryon-drag epochs and Hubble parameter, it provides a standard ruler with which to measure the late time evolution of the universe. This will be expanded on later.

After recombination, the photons are able to decouple and free-stream away from the baryons. Lacking any coupling, the baryons then begin to evolve purely under the effect of the gravitational potential sourced by the combination of the the dark matter and baryonic densities. Due to the redshifting of radiation and the limits enforced by

the Meszaros equation, matter has come to dominate long before recombination and the contribution to the overall density from radiation is negligible. Additionally, although the baryons introduce some small pressure, the non-interacting nature of the CDM means that $P_m << \rho_m$. As such the Continuity, Euler and Poisson equations can be written

$$\delta'_m + \nabla \cdot \boldsymbol{v}_m = -3\Phi', \tag{1.66}$$

$$\boldsymbol{v}'_m + \mathcal{H}\boldsymbol{v}_m + c_s^2 \nabla \delta_m = -\nabla\Phi, \tag{1.67}$$

$$\frac{3}{2}\Omega_m(\eta)\mathcal{H}^2\delta_m = \nabla^2\Phi. \tag{1.68}$$

These are very similar to the equations for CDM perturbations during the pre-recombination epoch, however now they apply to the combined CDM and baryonic properties and the baryonic pressure introduces a sound speed into the Euler equation. These could be neglected previously because the photons effectively stopped the baryons from contributing prior to recombination.

In the same way as previously, utilising a Fourier transformation, these can be combined into a single expression in the Newtonian regime

$$\delta''_m + \mathcal{H}\delta'_m + \left(k^2 c_s^2 - \frac{3}{2}\Omega_m(\eta)\mathcal{H}^2\right)\delta_m = 0. \tag{1.69}$$

For a given ratio of baryons and dark matter, this equation has two different solutions depending on the scale of the fluctuation, with a critical transition occuring at the Jeans' wavenumber

$$k_j = \frac{\mathcal{H}}{c_s}\sqrt{\frac{3}{2}\Omega_m(\eta)}. \tag{1.70}$$

On small scales where $k > k_j$, the equation has oscillatory solutions, as the baryonic pressure counteracts the gravitational potential from the combined matter. On large scales the gravitational potential 'wins' and the density has growing and decaying modes, the former arising from the balance between gravitational collapse and Hubble expansion and the latter quickly dissipating, such that only the growing mode is important. In the case where the matter density can be approximated as the CDM density then this solution is exactly the same as the $y >> 1$ solution to the Meszaros equation. The exact expression for this growing mode is given further on in this section.

Observations of the low redshift ($z < 3$) universe show that its evolution is now dominated by the cosmological constant. Solutions for the evolution of matter perturbations in this regime can be solved using the same method as for the matter perturbations in the post-recombination epoch (or equivalently the $y >> 1$ Meszaros equation). Indeed, because radiation still does not contribute to the gravitational potential the large scale limit of Eq 1.69 is still applicable, but the conformal Hubble parameter and matter density $\Omega_m(\eta)$ must be modified appropriately to account for the extra energy density from the cosmological constant.

1.2.5 Transfer Function and Power Spectrum

So far this section has presented a detailed analysis of the growth of density perturbations from the initial seeds left by quantum fluctuations during a period of rapid inflation, up to the Λ dominated regime after recombination. Throughout this description of the evolution of perturbations throughout cosmic time, it has been assumed that the perturbations are well within the horizon such that we can neglect the time derivatives in the gravitational potential. In reality the evolution of perturbations prior to recombination is a complex function of the scale of the perturbations and when those scales enter the Hubble volume, making the transition from superhorizon to subhorizon modes. For a given k mode the exact evolution hence depends on solutions to the relativistic Boltzmann equations at some early time which connects to the subhorizon Newtonian solutions at late times. The full set of these solutions up to the point at which all modes of interest are within the horizon is called the Transfer Function, $T(k)$.

In linear theory, at late times, all modes of interest have already entered the horizon and as such all modes evolve in the same way, there is no scale dependence. Prior to this there is no strict time dependence for the solutions as each mode will enter the horizon at a specific time, such that the transfer function for each individual mode will encapsulate the time-dependence automatically. As such the transfer function is only a function of scale.

This then presents a convenient way of decomposing the evolution of the primordial perturbations in the early universe into a scale dependent part, and a growth function dependent on the scale factor. An applicable growth function will be presented shortly in the form of the linear growth factor $D_1(a)$. Hence a linear primordial density fluctuation $\delta_m(\boldsymbol{k}, 0)$ becomes a late time linear density fluctuation via

$$\delta_m(\boldsymbol{k}, a) = T(\boldsymbol{k}) D_1(a) \delta_m(\boldsymbol{k}, 0). \tag{1.71}$$

If the primordial density perturbations are assumed to be approximately Gaussian, then, much like with the CMB anisotropies, the statistic of interest is the power spectrum of the late time density perturbations. This is defined as

$$\langle \delta_m(\boldsymbol{k}, a) \delta_m^*(\boldsymbol{k}', a) \rangle = (2\pi)^3 P_m(k, a) \delta^D(\boldsymbol{k} - \boldsymbol{k}'), \tag{1.72}$$

where $\delta^D(\boldsymbol{k} - \boldsymbol{k}')$ is the Dirac delta function. The relation between the late time density perturbations and the primordial perturbations in turn means that the late time matter power spectrum can be written

$$P_m(k, a) = T^2(k) D_1^2(a) P_m(k, 0) \tag{1.73}$$

For primordial density fluctuations it is often assumed that

$$P_m(k, 0) \propto k^{n_s} \tag{1.74}$$

and the scale dependence $n_s \approx 1$. In truth this value is measured to be $n_s = 0.9667 \pm 0.0040$ (Planck Collaboration et al., 2015a), which is very close to the exact value predicted by theories of inflation.

The amplitude of the primordial power spectrum can be measured using the anisotropies in the CMB. From there a value for the amplitude of the power spectrum at late times can be inferred, under the assumption of a model for gravity. By convention, the clustering ampltiude at late times is defined as the root-mean-squared density fluctuation in spheres of radius $r = 8\,h^{-1}\,\mathrm{Mpc}$, σ_8. This is related to the power spectrum via

$$\sigma_r^2 = \frac{1}{2\pi^2}\int_0^\infty P_m(k)W^2(k,r)k^2dk, \tag{1.75}$$

where $W(k,r)$ is the fourier transform of spherical top-hat filter of radius r.

$$W(k,r) = 3\left(\frac{\sin(kr)}{k^3r^3} - \frac{\cos(kr)}{k^2r^2}\right) \tag{1.76}$$

Codes such as CAMB (Lewis et al., 2000) which evaluate the full set of Einstein Field and Boltzmann equations can be used to evaluate the linear matter transfer function and power spectrum in addition to the CMB power spectrum. Similarly the growth function can also modelled numerically. However, the true cosmology of our universe is unknown. To identify the underlying cosmology of our universe, in terms of its matter density etc., the late-time power spectrum can be measured and compared to that predicted from Eq. 1.73.

1.2.6 Linear Growth

Solutions to the matter equation of motion (Eq. 1.69) above the Jean's wavelength give rise to a dominant growing mode balancing the attractive force of gravity and the Hubble expansion. In the linear regime this solution takes the form of Eq. 1.71. Heath (1977) showed that for any flat cosmology with arbitrary matter, radiation and cosmological constant energy density, the linear growth factor (at a given scale factor) can be written as

$$D_1(a) = AH(a)\int_0^a \frac{da'}{(a'H(a'))^3}. \tag{1.77}$$

The constant of proportionality A can be derived by requiring that during matter domination the previous solutions covered in this chapter are recovered, $D_1(a) = a$. At this time $H(a) = H_0\sqrt{\Omega_m(1)/a^3}$ and as such

$$D_1(a) = \frac{5\Omega_{m,0}}{2}\frac{H(a)}{H_0}\int_0^a \frac{da'}{(a'H(a')/H_0)^3}. \tag{1.78}$$

This is an extremely powerful solution, and allows one to express the present day clustering of matter on linear scales to the fluctuations shortly after the epoch of recombination. A similarly powerful quantity is the linear growth rate, which is the logarithmic

differential of the linear growth factor,

$$f = \frac{a}{D_1}\frac{dD_1}{da}. \tag{1.79}$$

Using the linear growth rate, one can relate the infall velocity towards a density perturbation to the amplitude of the perturbation. In the linear regime, on sub-horizon scales, the k-space continuity equation (Eq. 1.37) for the matter states

$$\boldsymbol{v}(\boldsymbol{k}) = -\frac{i}{k}\delta'(\boldsymbol{k})\hat{\boldsymbol{k}}. \tag{1.80}$$

But the late time evolution of the overdensity is well known on linear scales, it scales with respect to the linear growth factor. As such

$$\boldsymbol{v}(\boldsymbol{k}) = -\frac{i}{k}\frac{\delta(\boldsymbol{k})}{D_1}D_1'\hat{\boldsymbol{k}}. \tag{1.81}$$

Converting from conformal time to scale factor,

$$\boldsymbol{v}(\boldsymbol{k}) = -\frac{ifaH\delta(\boldsymbol{k})}{k}\hat{\boldsymbol{k}}. \tag{1.82}$$

The relation between the velocity and density fields in terms on the growth rate depends fundamentally on the theory of GR. As such the linear growth rate is a key observable in the large scale structure of the universe, as it allows constraints to be put on the fidelity of General Relativity. There are several ways to parameterise f to test GR. Linder & Cahn (2007) advocate the simple form,

$$f(a) \approx \Omega_m(a)^{\gamma} \tag{1.83}$$

where from General Relativity, the growth index $\gamma = 0.554$. Measuring this γ parameter is one of the simplest ways of providing a consistency test of GR and can be used to test other theories of modified gravity. Theories of modified gravity are a very active area of research and provide an alternative to the idea of dark matter. In modified gravity theories, rather than an 'invisible' sector of matter contributing to the gravitational potential, the form of the gravitational potential from GR is instead incorrect on large scales.

To date, the most robust way of measuring f and hence γ is using Redshift Space Distortions in the two-point clustering of matter. The basic theory behind this signal is given in Section 1.4. An example of how RSD can be used on data to test the consistency of GR is the focus of Chapter 4. The use of RSD is just one technique that is used to investigate the correctness of the assumptions that go into the Hot Big Bang Model and constrain its unknown parameters. The next section gives an overview of a variety of probes of the cosmology of the universe and the constraints from these that led to the Hot Big Bang model becoming the concordance model.

1.3 Observational Probes of Cosmology

So far this chapter has given a detailed introduction to the Hot Big Bang model. Within the framework of GR, the FLRW metric and the assumption of a ΛCDM model for the energy and momentum, the evolution of perturbations from inflation till the present day has been presented. However it is only recently that this picture has emerged as a concordance model due to a large number of corroborating, but independent, observations. Though the Hot Big Bang model is widely accepted and has been for some time, there are still many questions to which the answer is unknown:

Is there Cold Dark Matter? - Still no candidate particle for dark matter has been found. This has led many to believe that there is no dark matter, and instead it is the assumption of General Relativity that is incorrect. On solar system scales there can be no doubt that General Relativity reproduces observations such as the precision of the orbit of Mercury with astonishing accuracy yet the constraints on larger scales are much weaker. Only by performing more stringent tests of GR on cosmological scales can this be addressed. There are a wealth of alternate modified theories of gravity that use the introduction of massive force carriers for gravity or other additional degrees of freedom that could provide the solution, but many of these remain untested.

What is Dark Energy? - Similarly, the fact that the late-time universe is undergoing accelerated expansion has been so well established in recent years by observations, that this phenomenon has been adopted into the concordance cosmological model. However the exact driving force behind this accelerated expansion is unknown. Current constraints point towards a cosmological constant Λ, with density Ω_Λ and equation of state $w = -1$. However these constraints still leave significant room for some other equation of state and hence some other cause for dark energy. Modified theories of gravity can also predict accelerated expansion but these too remain largely untested. It could be that again the assumption of GR is to blame or there is some other field causing the acceleration.

What is the nature of Inflation? - As with dark energy, inflation has been adopted into the concordance model, largely due to its ability to predict the uniformity of the CMB and solve the Horizon, Flatness and Relic problems. The form of the Inflaton that causes inflation is still poorly understood however. Many models of inflation predict small non-Gaussian contributions to the initial density field which can be detected post-recombination and could shine a light on this question.

What are the abundance of matter, radiation and dark energy? - Even if GR and ΛCDM provide the correct prescription for the universe, there are still several unknown parameters in these theories. The relative abundances of the components of the energy-momentum tensor have a large effect on the evolution of the universe. Recent probes have put tight constraints on these parameters, yet degeneracies between them mean that more

can still be done to pinpoint their exact values within the concordance model.

Although there are still many unknowns, the current state of the concordance model is very different from 50 years ago. This is due to the overwhelming number of observations of the universe made in that time. In the rest of this section these probes will be presented individually and their contribution to the answers to the above questions will be given.

1.3.1 CMB

The Cosmic Microwave Background gives the most robust constraints on the ΛCDM cosmological parameters from any single probe. A detailed report on the Cosmic Microwave Background was given in Section 1.2.4. Here the state-of-the-art measurements will be presented. Modelling of the CMB anisotropy power spectra from the Planck satellite alone gives the extremely tight measurements $\Omega_b h^2 = 0.02225 \pm 0.00016$ and $\Omega_c h^2 = 0.1198 \pm 0.0015$ (Planck Collaboration et al., 2015a). Strong constraints on the scalar index of the primordial power spectrum are obtained, $n_s = 0.9645 \pm 0.0049$, which helps to limit the allowed form of Inflation. Assuming ΛCDM, strong constraints on the Hubble Parameter, $H_0 = 67.27 \pm 0.66\,\mathrm{kms}^{-1}\mathrm{Mpc}^{-1}$, and power spectrum variance, $\sigma_8 = 0.831 \pm 0.013$, are also obtained. Differences between these values and measurements that do not require the assumption of ΛCDM/GR could provide evidence that these assumptions are incorrect on some scales.

The CMB also provides the strongest constraints on the amount of non-Gaussianity in the primordial perturbations arising from inflation, hence constraining the form of the Inflaton and models of Inflation. Primordial non-Gaussianity gives rise to an extra signal in the three-point clustering of the CMB anisotropies, on top of that caused by processes between the observer and the surface of last scattering. Different triangular shapes in the primordial three-point 'bispectrum' correspond to different models of Inflation, and can be split into three categories, 'squeezed' or 'local' configurations, 'equilateral' triangles and 'folded' or 'orthogonal' configurations. Planck Collaboration et al. (2015b) give very tight constraints on the amplitude of the non-Gaussian bispectrum in these configurations: $f_{NL}^{local} = 0.8 \pm 5.0$, $f_{NL}^{equil} = -4 \pm 43$ and $f_{NL}^{ortho} = -26 \pm 21$, consistent with zero. These low values put severe restrictions on the allowed models of inflation.

1.3.2 Supernovae Type IA

Due to the nature of their formation, Type 1A Supernovae can be used as standard candles, i.e. their luminosity distance can be inferred purely from their apparent luminosity, without the need for an assumption of the cosmological model. Type 1A Supernovae are

formed when a white dwarf accretes matter from a companion star. When it reaches the Chandrasekhar limit (Chandrasekhar, 1931) it undergoes a cataclysmic supernova explosion. Because of the exact mass limit at which collapse occurs, the instrinsic luminosity of the supernova can be inferred from the shape of its 'light curve', the measured apparent magnitude as a function of time (Pskovskii, 1977; Phillips, 1993).

It was using the properties of Type 1A supernovae that Perlmutter et al. (1999) and Riess et al. (1998) first detected the accelerated expansion of the universe, through measurements of the Hubble Parameter. Today Type 1A supernovae help provide some of the strongest independent constraints on Ω_m and H_0. Betoule et al. (2014) find an independent value of $\Omega_m = 0.295 \pm 0.024$, which is consistent with the latest CMB measurements It is their data that is used to constrain the dark energy equation of state in Figure 1.4. Riess et al. (2011) find constraints on the Hubble parameter of $H_0 = 73.8 \pm 2.4\,\mathrm{kms}^{-1}\mathrm{Mpc}^{-1}$ This is in tension with the CMB results of Planck Collaboration et al. (2015a) by approximately 2.5σ, however there is equally significant variation in the values found from different supernovae samples (Freedman et al., 2012; Tammann & Reindl, 2013) and CMB measurements Bennett et al. (2013). As such it is more likely that this tension is caused by systematic errors within the different analyses that have not been accounted for as opposed to physics beyond the concordance model. More work is needed to pinpoint the source of this discrepancy.

1.3.3 Lensing

Light from distant galaxies can be distorted by intervening matter as it travels towards the observer. Light rays travel on null geodesics and as such if the light travels though a gravitational potential the path the light travels can be bent, resulting in lensing of the object from which the light originates. As this depends on only the gravitational potential, it offers a much more direct route for testing GR than other probes such as RSD. For light travelling from distant galaxies through the most massive objects, the lensing can be visually apparent, as the light received by the observer is so distorted as to make the host galaxy appear as a ring or arc around the foreground object. This is called strong lensing. Such events are rare however. Weak lensing on the other hand, a distortion in the galaxy shape on the order of $\sim 1\%$, is a much more subtle effect but the number of objects that undergo some sort of weak lensing is much higher.

The effect of lensing is to turn circular objects on the sky into elliptical ones, an effect of the 'shear' created by the gravitational potential, and to magnify them. However the fact that most galaxies are intrinsically elliptical creates difficulty in measuring the weak lensing on an individual basis. Instead the gravitational shear is measured by computing the two-point shear clustering, averaging over many objects. The shear power

spectrum is sensitive to the comoving distances to the objects and to the matter power spectra, averaged along the line of sight. As such it provides constraints on Ω_m, σ_8 and w although as with other probes these are degenerate and can only be measured individually in combination with other data. The most precise current constraints come from the CFHTLens survey (Heymans et al., 2012) which gives the combined measurement $\sigma_8(\Omega_m/0.27)^{0.6} = 0.79 \pm 0.03$. Using the value of Ω_m from Planck Collaboration et al. (2015a) shows that there is some tension between this constraint and the value Planck Collaboration et al. (2015a) themselves obtain assuming GR. Weak lensing is a relatively new technique and modelling of the galaxy shapes and intrinsic shape alignments between different galaxies is a difficult task, so whether this discrepancy is an indication of new physics or unknown systematics remains to be seen.

1.3.4 Clusters of Galaxies

Clusters of galaxies are the largest virialised objects in the universe and they form in the most massive dark matter halos. Although they are formed via complex non-linear and astrophysical processes, the large size of these objects allows them to bridge the transition region between the linear and non-linear scales. In particular the number density of galaxy clusters as a function of mass depends strongly on the matter power spectrum, and as such Ω_m and σ_8. w can also be constrained by looking at the evolution of the mass function with redshift and combining with CMB and BAO datasets. Henry et al. (2009); Vikhlinin et al. (2009); Mantz et al. (2010); Rozo et al. (2010) all present constraints on Ω_m and σ_8 using the cluster mass function which are within reasonable agreement with the results of Planck Collaboration et al. (2015a), providing evidence for the fidelity of GR.

Additional information on modified theories of gravity can be garnered from galaxy clusters by using a combination of lensing and the x-ray luminosities measured from the Intergalactic Medium. The large size of clusters make them prime candidates for investigating the screening mechanism of 'chameleon' gravity theories (Waterhouse, 2006), where the modification to GR disappears in high density environments (i.e., within the cluster to match solar system constraints) but becomes evident on large scales outside the cluster virial radius. Recently Terukina et al. (2014) and Wilcox et al. (2015) used clusters to place constraints on the strength of the modification to GR in chameleon gravity.

1.3.5 Large Scale Structure

The large scale structure of galaxies across the universe allows for direct measurements of the galaxy power spectrum and hence for inference of the underlying matter power spectrum. In fact there is a wealth of information that can be gathered based on the galaxy clustering. Two of the main features in the clustering are the BAO, detailed in

Section 1.2.4, which provides a standard ruler on the sky for measuring Ω_m, H_0 and w, and RSD which allows for tests of GR and modified gravity. Measurements of these probes form a large part of this thesis and so they will be discussed in more detail in the next section.

There are also other features of interest in the galaxy power spectrum. Primoridial non-Gaussianity can introduce scale-dependent effects on the largest scales of the galaxy power spectrum (Dalal et al., 2008). Such an effect was measured by Ross et al. (2013) who found $-45 < f_{NL}^{local} < 195$, which whilst not competitve compared to the constraints from Planck Collaboration et al. (2015a), does provide an independent way of meauring this that will undoubtedly find use in future surveys.

1.3.6 Combined Probes

Like the Ω_m-σ_8 constraints from lensing data or the $\Omega_b h^2$ constraints from the CMB, most cosmological probes of the universe suffer from degeneracies between parameters which weaken the measurements on individual parameters. This is especially true when looking beyond the 'concordance' model and testing GR and ΛCDM. Information on, for example, the equation of state of dark energy or the growth index γ, can only be obtained when multiple probes are combined to break the degeneracies between parameters.

The strength of the CMB constraints is such that these are often the basis for the consensus cosmology parameters obtained from a range of probes. The current consensus values for the concordance cosmological model, from the combination of CMB, BAO and Supernovae data, are (Planck Collaboration et al., 2015a)

$$\begin{aligned} \Omega_{b,0}h^2 &= 0.02230 \pm 0.00014 \\ \Omega_{c,0}h^2 &= 0.1188 \pm 0.0010 \\ \Omega_{\Lambda,0} &= 0.6911 \pm 0.0062 \\ \Omega_{r,0}h^2 &= 4.18343 \times 10^{-5} \\ \Omega_{k,0} &= 8^{+40}_{-39} \times 10^{-4} \end{aligned} \tag{1.84}$$

The baryonic, cold dark matter and radiation densities are presented in the form in which the CMB constrains them. They are degenerate with the Hubble parameter. The value of $\Omega_{r,0}h^2$ is derived to high precision directly from the measured temperature of the CMB, $T = 2.7255 \pm 0.0006K$ (Mather et al., 1994).

Combining the CMB power spectrum with Supernovae, Lensing, Cluster and LSS data provides a route to obtaining tight constraints on extensions to these models such as the equation of state of dark energy. Figure 1.4 (taken from Planck Collaboration et al. 2015a) shows constraints on the time-varying equation of state of dark energy from

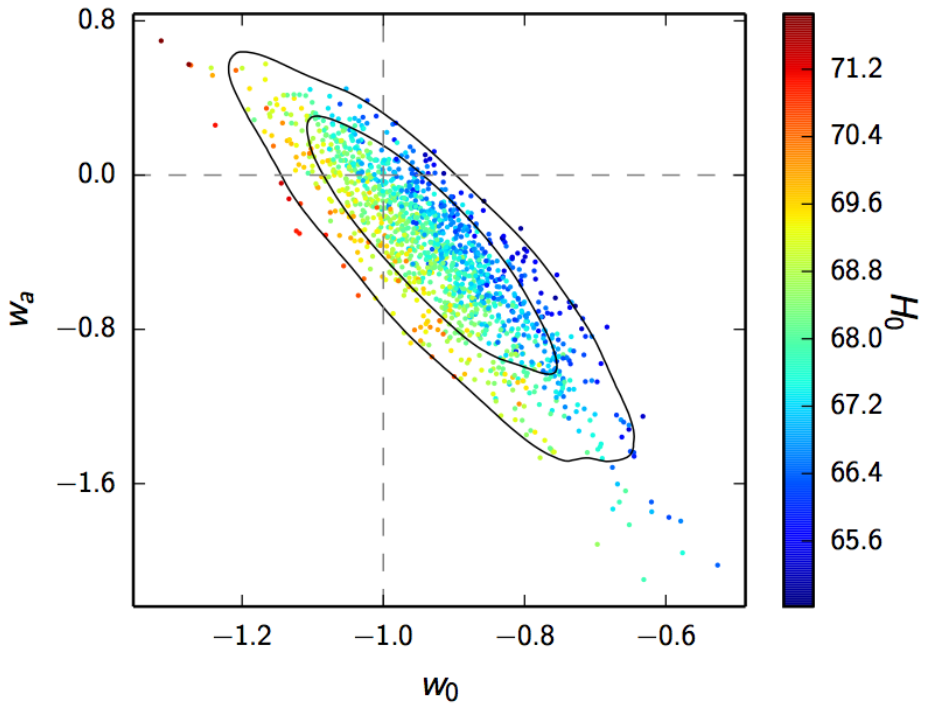


Figure 1.4: Constraints on the time-varying dark energy equation of state from Planck Collaboration et al. (2015a) using the combination of CMB, BAO and Type 1A supernovae data. The points show individual links in the likelihood chains, whilst the black lines show the 1 and 2σ contours. The dashed lines show the ΛCDM prediction, $w_0 = -1, w_a = 0$.

this combination of data Supernovae, BAO and CMB data, under the assumption of an equation of state given by $w = w_0 + (1 - a)w_a$.

This combination of data is in good agreement with the predictions of the ΛCDM model, however the constraints still leave much room for non-standard dark energy models, with different equations of state. Only future datasets have the potential to improve these constraints and pin down the exact nature of the accelerated expansion of the universe.

Similarly, the combined probes of Weak Lensing, LSS (namely BAO and RSD) and the CMB can be used to constrain the growth rate and test GR. Figure 1.5 shows the joint constraints obtained from BAO and RSD analysis of data from the Baryon Oscillation Spectroscopic Survey, the Planck satellite, Type 1A Supernovae probes and measurements of the Hubble constant through the local distance ladder. Published in Samushia et al. (2014) this shows that this combination of data gives a growth index in agreement with the predictions from GR, although with a slight preference for weaker gravitational interactions. Accurate analysis of larger datasets in the future will shrink these contours,

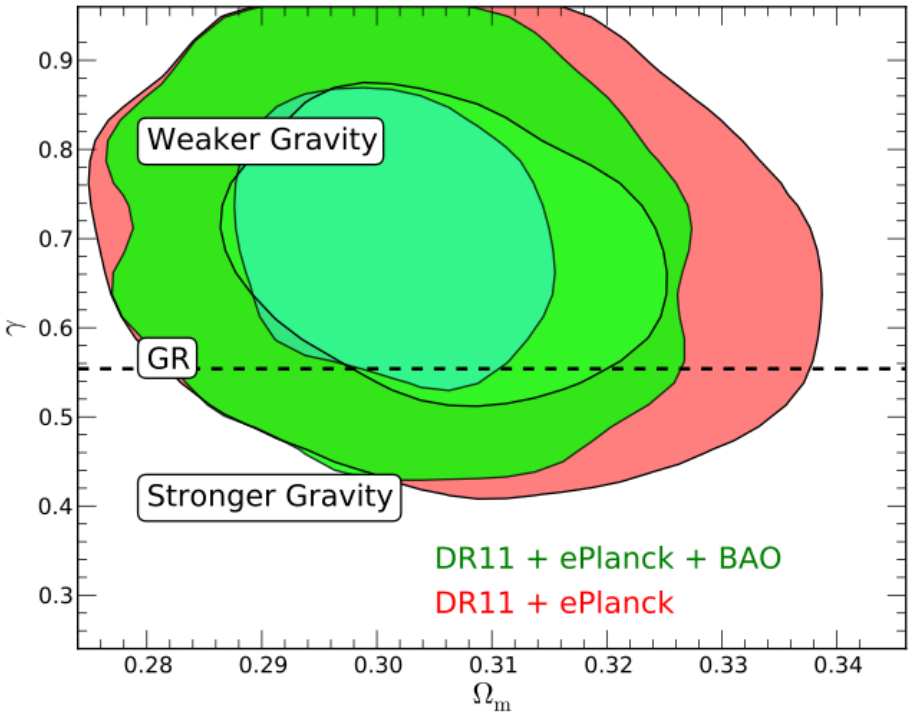


Figure 1.5: Constraints on the growth index from Samushia et al. (2014) using the combination of CMB, RSD, BAO, Type 1A supernovae and local H_0 data. The contours show the 1 and 2σ constraints with and without the inclusion of anisotropic BAO information. The dashed line is the prediction of GR, which agrees with the constraints, although there is some preference for models with weaker gravitational interactions.

and if the observed trend remains, provide evidence for the inaccuracy of GR on large scales.

1.4 Measuring Large Scale Structure

As seen in Section 1.2.5, the matter power spectrum holds a wealth of information. In fact, for perturbations drawn from a multivariate Gaussian density field with $\langle \delta_m \rangle = 0$, the matter power spectrum and its associated covariance matrix, hold *all* the information about the density field. Although the non-linear nature of gravitational collapse actually introduces some non-Gaussianity into the late-time universe, as can the presence of primordial non-Gaussianity during inflation, the non-Gaussian contributions to the underlying distribution of the density perturbations are small, and the power spectrum still contains the majority of the cosmological information.

Measurements of the clustering of the large scale structure (LSS) of the universe currently provide the most robust route to measuring the late time matter power spectrum

and hence the cosmology dependence of the growth of structure. Typically, probes of LSS use measurements of the clustering of large numbers of galaxies throughout the universe to probe cosmology. Alongside the overall shape of the power spectrum, there are two main features within the galaxy distribution that allow characterisation of the underlying cosmology of the universe, the angular scale of Baryon Acoustic Oscillations compared to their sound horizon at the baryon-drag epoch, and Redshift Space Distortions caused by the infalling of galaxies into gravitational wells. How these are used for cosmological constraints will be detailed in this section. First though, an overview of how the 3D distribution of galaxies can be statistically analysed and the matter power spectrum measured will be presented.

1.4.1 Characterising the Galaxy Overdensity Field

Much like the temperature anisotropies in the CMB tracing the density fluctuations in the early universe, the density of structure in the late-time universe can be measured using large numbers of galaxies. The galaxy overdensity $\delta_g(\boldsymbol{r})$ at some position $\boldsymbol{r}$ can be calculated by counting the number density of galaxies at that position and comparing this to the expected number of galaxies,

$$\delta_g(\boldsymbol{r}) = \frac{\rho_g(\boldsymbol{r}) - \bar{\rho_g}(\boldsymbol{r})}{\bar{\rho_g}(\boldsymbol{r})}. \tag{1.85}$$

However the quantity of interest is the overdensity of matter at a given location, as opposed to the overdensity of galaxies. Although the luminous galaxies in the universe trace the underlying matter distribution, there is not a one-to-one correlation. This means that the galaxy overdensity field is related by some bias function to the matter overdensity. This bias function is nominally a complex, non-linear function of galaxy properties such as type, color, morphology and environment as well as the scale of the overdensity. However in the linear regime, for a given measured set of galaxies, this is often approximated as a simple multiplicative constant

$$\delta_g(\boldsymbol{r}) = b\delta_m(\boldsymbol{r}). \tag{1.86}$$

Under some bias model, it is possible to define the correlations between matter overdensities at different locations using the galaxies. In general this leads to a series of n-point correlation functions, the first of which is the two-point correlation function, the Fourier transform of the power spectrum. The next order correlation function, the three-point has the bispectrum as its Fourier transform, which, in this work, will be touched on only briefly in Chapter 2 and so will not be presented in detail here.

1.4.2 Two-point Clustering

Two-point Correlation Function

The probability of finding a galaxy in an infinitesimally small volume element δV can be written

$$\delta P = \bar{n}(\boldsymbol{r})\delta V \tag{1.87}$$

where $\bar{n}$ is the mean number density of galaxies.

The two-point correlation function, ξ, is then defined using the probability of finding two galaxies in the universe separated by some vector $\boldsymbol{r}_{12} = \boldsymbol{r}_1 - \boldsymbol{r}_2$,

$$\delta P = \bar{n}^2 \delta V_1 \delta V_2 (1 + \xi_g(\boldsymbol{r}_{12})). \tag{1.88}$$

For a completely unclustered collection of galaxies $\xi_g(\boldsymbol{r}_{12}) = 0$. Choosing the first galaxy at random, the probability of finding a neighbour separated by a vector $\boldsymbol{r}'$ is then, in terms of the galaxy overdensity,

$$\xi_g(\boldsymbol{r}') = \langle \delta_g(\boldsymbol{r} + \boldsymbol{r}')\delta_g(\boldsymbol{r}) \rangle. \tag{1.89}$$

This expression shows the two-point correlation function to be the Fourier transform of the Power spectrum. The optimal method of estimating this from a galaxy field based on the probabilistic definition above is given by Landy & Szalay (1993)

$$\xi_g(r, \mu) = \frac{DD(r, \mu) - 2DR(r, \mu) + RR(r, \mu)}{RR(r, \mu)}. \tag{1.90}$$

Here the separation vector $\boldsymbol{r}'$ has been split into length, r, and angular, μ components. A field of random points is used to estimate the expected mean galaxy density at a given location and hence the expected number of pairs separated by r andμ. DD corresponds to the normalised number of galaxy pairs for a given separation, DR the number of galaxy-random pairs and RR the number of random-random pairs.

Power spectrum

The correlation function and Power spectrum form a Fourier pair,

$$\xi(\boldsymbol{r}) = \int \frac{d^3\boldsymbol{k}}{(2\pi)^3} P(\boldsymbol{k}) e^{-i\boldsymbol{k}\cdot\boldsymbol{r}}. \tag{1.91}$$

Feldman et al. (1994) provide a well known estimator (the FKP estimator) for the power spectrum of a weighted galaxy overdensity field using this property. A detailed derivation of this estimator will be provided in Chapter 5. For now a quick pedagogical overview will suffice. The estimator starts with a weighted galaxy overdensity, with a suitable normalisation A,

$$F(\boldsymbol{r}) = \frac{w(\boldsymbol{r})\bar{n}(\boldsymbol{r})\delta_g(\boldsymbol{r})}{A}. \tag{1.92}$$

The ensemble average of the square of this function for different $\boldsymbol{r}$ gives the weighted galaxy correlation function. If instead it is Fourier transformed and the ensemble average taken

$$\langle |F(\boldsymbol{k})|^2 \rangle \approx P_g(\boldsymbol{k}) + P_{shot} \tag{1.93}$$

P_{shot} is the residual *shot-noise* component arising from the fact that the density field is a discrete sample of galaxies, as opposed to a continuous distribution. From this the estimator of the power spectrum, averaged over some shell in k-space of volume V_k, can be written as

$$\hat{P}_g(k) = \int_{V_k} \frac{d^3\boldsymbol{k}'}{V_k} |F(\boldsymbol{k}')|^2 - P_{shot} \tag{1.94}$$

In practice, implementing this expression is done by converting the integral into a sum over the N_k k-modes within the bin in question, and approximating $F(\boldsymbol{k})$ and P_{shot} via a sum over galaxies, g, and random points, s, much like the two-point correlation function estimator of Landy & Szalay (1993),

$$\hat{P}_g(k) = \frac{1}{N_k} \sum_{k<|\boldsymbol{k}|<k+\delta k} [|F(\boldsymbol{k})|^2 - P_{shot}] \tag{1.95}$$

where

$$F(\boldsymbol{k}) = \sum_g w(\boldsymbol{r}_g) e^{i\boldsymbol{k}\cdot\boldsymbol{r}_g} - \alpha \sum_s w(\boldsymbol{r}_s) e^{i\boldsymbol{k}\cdot\boldsymbol{r}_s} \tag{1.96}$$

$$P_{shot} = \alpha(1+\alpha) \sum_s w^2(\boldsymbol{r}_s) e^{i\boldsymbol{k}\cdot\boldsymbol{r}_s} \tag{1.97}$$

and α is the ratio of the number of galaxies to randoms points used. Feldman et al. (1994) find that the minimum variance weight for this estimator is

$$w(\boldsymbol{r}) = \frac{1}{1 + \bar{n}(\boldsymbol{r}) P_g(k)}, \tag{1.98}$$

which is named the FKP weight after their seminal paper.

Covariance Matrix and Effective Volume

In the same work, Feldman et al. (1994) also investigate the covariance matrix of the power spectrum, and the variance in their estimator. Again a detailed derivation of the covariance matrix will be presented in Chapter 5. In brief, the variance in the FKP estimator of the power spectrum is

$$\sigma_p^2(k) = \int_{V_k} \frac{d^3\boldsymbol{k}'}{V_k} \int_{V_k} \frac{d^3\boldsymbol{k}''}{V_k} \langle \hat{P}_g(\boldsymbol{k}') \hat{P}_g(\boldsymbol{k}'') \rangle - \langle \hat{P}_g(\boldsymbol{k}') \rangle \langle \hat{P}_g(\boldsymbol{k}'') \rangle \tag{1.99}$$

Assuming the galaxy overdensity field is drawn from a Gaussian distribution, Feldman et al. (1994) show that

$$\langle \hat{P}_g(\boldsymbol{k}') \hat{P}_g(\boldsymbol{k}'') \rangle - \langle \hat{P}_g(\boldsymbol{k}') \rangle \langle \hat{P}_g(\boldsymbol{k}'') \rangle = |\langle F(\boldsymbol{k}') F^*(\boldsymbol{k}'') \rangle|^2. \tag{1.100}$$

Using the same method as to derive the power spectrum estimator in the first place, the variance in the power spectrum estimator can be written

$$\sigma_p^2(k) = \frac{(2\pi)^3 P_g^2(k)}{V_k} \frac{\int d^3\boldsymbol{r} \bar{n}^4(\boldsymbol{r}) w^4(\boldsymbol{r}) [1 + 1/\bar{n}(\boldsymbol{r}) P_g(k)]^2}{\left[\int d^3\boldsymbol{r} \bar{n}^2(\boldsymbol{r}) w^2(\boldsymbol{r})\right]^2}. \tag{1.101}$$

This expression was recomputed by Tegmark (1997) to show that the covariance matrix for the power spectrum, in the absence of a survey window function and assuming Gaussianity, consists of only diagonal elements given by

$$C(k_i, k_j) = \frac{2P_g(k_i)P_g(k_j)}{V_n(k_i)V_{eff}(k_i)} \delta^D(k_i - k_j), \tag{1.102}$$

where $V_n(k_i) = 4\pi k_i^2 \Delta k_i/(2\pi)^3$ is the k-space volume of a thin shell of width Δk_i centred at k_i and

$$V_{eff} = \int d^3\boldsymbol{r} \left[\frac{\bar{n}(\boldsymbol{r})P_g(k)}{1 + \bar{n}(\boldsymbol{r})P_g(k)}\right]^2 \tag{1.103}$$

is the effective volume. This approximation for the covariance matrix is often used as the basis for Fisher matrix forecasts of constraints on cosmological surveys. The effective volume is also used to determine the effective redshift, z_{eff}, of LSS measurements,

$$z_{eff} = \frac{\int z dV_{eff}}{\int dV_{eff}} = \frac{\int_{z_1}^{z_2} \frac{z}{H(z)} D_c^2(z) \left(\frac{\bar{n}(z)}{1+\bar{n}(z)P_g(k)}\right)^2 dz}{\int_{z_1}^{z_2} \frac{1}{H(z)} D_c^2(z) \left(\frac{\bar{n}(z)}{1+\bar{n}(z)P_g(k)}\right)^2 dz}. \tag{1.104}$$

In the last equality, Eq. 1.103 has been substituted in, the 3D integral converted to distinct integrals over the angle and comoving distance, and Eq. 1.26 used to relate the comoving distance to the redshift.

Ideally, cosmological measurements from LSS would be obtained for infinitesimally narrow redshift bins, such that the time evolution of the cosmology of the universe can be captured. In reality the number density of galaxies in the universe, in combination with the signal-to-noise of a given experiment, means that measurements of galaxies across some range of redshifts must be combined to obtain constraints. The effective redshift is the optimal redshift at which a measurement within some redshift bin, i.e., combining galaxies distributed between two different redshifts, should be quoted.

1.4.3 Baryon Acoustic Oscillations as a Standard Ruler

As derived in Sections 1.2.3 and 1.2.4, the acoustic oscillations in the photon-baryon plasma prior to recombination leave a distinct impression on the overdensity of baryons after recombination. Even though this baryonic overdensity at the sound horizon of the baryon-drag epoch is subdominant compared to the CDM overdensity, which pulls the baryonic material into the centres of potential wells, this imprint is still present, and

measurable, in the present day galaxy distribution. The slight overdensity remaining in the galaxy distribution shows up as a peak in the two-point correlation function at $\sim 100\,h^{-1}\,\mathrm{Mpc}$, and oscillations in the power spectrum, which are most prominent between the scales $k = 0.02 - 0.3\,h\,\mathrm{Mpc}^{-1}$. Examples of BAO features in the galaxy two-point correlation function and power spectrum from the SDSS-III Baryon Oscillation Spectroscopic Survey (BOSS) Data Release 9 (DR9, Anderson et al. 2014a) are shown in Figure 1.6

The first conclusive ($> 3\sigma$) detection of the Baryon Acoustic Features in the two-point clustering of galaxies was presented, independently, by Eisenstein et al. (2005) and Cole et al. (2005) using data from the Sloan Digital Sky Survey (SDSS) and 2 Degree Field Survey (2dF) respectively. Prior to this studies by Einasto et al. (1997), Miller et al. (2001) and Percival et al. (2001) had given tentative hints of the presence of the BAO feature in the large scale distribution of galaxies. Since then, a number of studies (Hütsi, 2006; Percival et al., 2007; Okumura et al., 2008; Gaztañaga et al., 2009; Kazin et al., 2010; Percival et al., 2010; Beutler et al., 2011; Blake et al., 2011a; Seo et al., 2012; Padmanabhan et al., 2012; Anderson et al., 2014a,b; Kazin et al., 2014; Ross et al., 2015) have reported detections of the BAO feature in the clustering of galaxies and used this to constrain the cosmology of the universe. The most significant detection of the BAO signal to date in both the two-point correlation function and power spectrum was made using the BOSS-DR11 data by Anderson et al. (2014b), with detections at 8σ and 7σ respectively. This highly robust measurement also allows for very precise measurements of the expansion rate of the universe and equation of state of dark energy at $z = 0.57$, using the BAO feature as a standard ruler.

As explained previously, the radius of the sound horizon at the baryon drag epoch sets up a characteristic scale at which there is an overdensity of baryonic material. Due to the nature of the photon-baryon plasma prior to recombination, the radius of the baryonic sound horizon at the baryon drag epoch is directly related to the scale of the anisotropies in the CMB, via the projected sound horizon of these anisotropies on the surface of last scattering. As the scale of the sound horizon, $r_d \approx 150\mathrm{Mpc}$, is large, the size of the BAO feature since the baryon drag epoch, in both the angular and radial directions, changes mainly due to the late time expansion of the universe, and so provides a method of determining the underlying cosmology determining the expansion rate. In particular, the angular size subtended by the BAO feature on the sky, $\Delta\theta$, is related to the angular diameter distance and the sound horizon by

$$D_A(z) = \frac{r_d}{\Delta\theta(1+z)}, \tag{1.105}$$

whilst the change in redshift of the BAO feature along the line of sight, relates to the

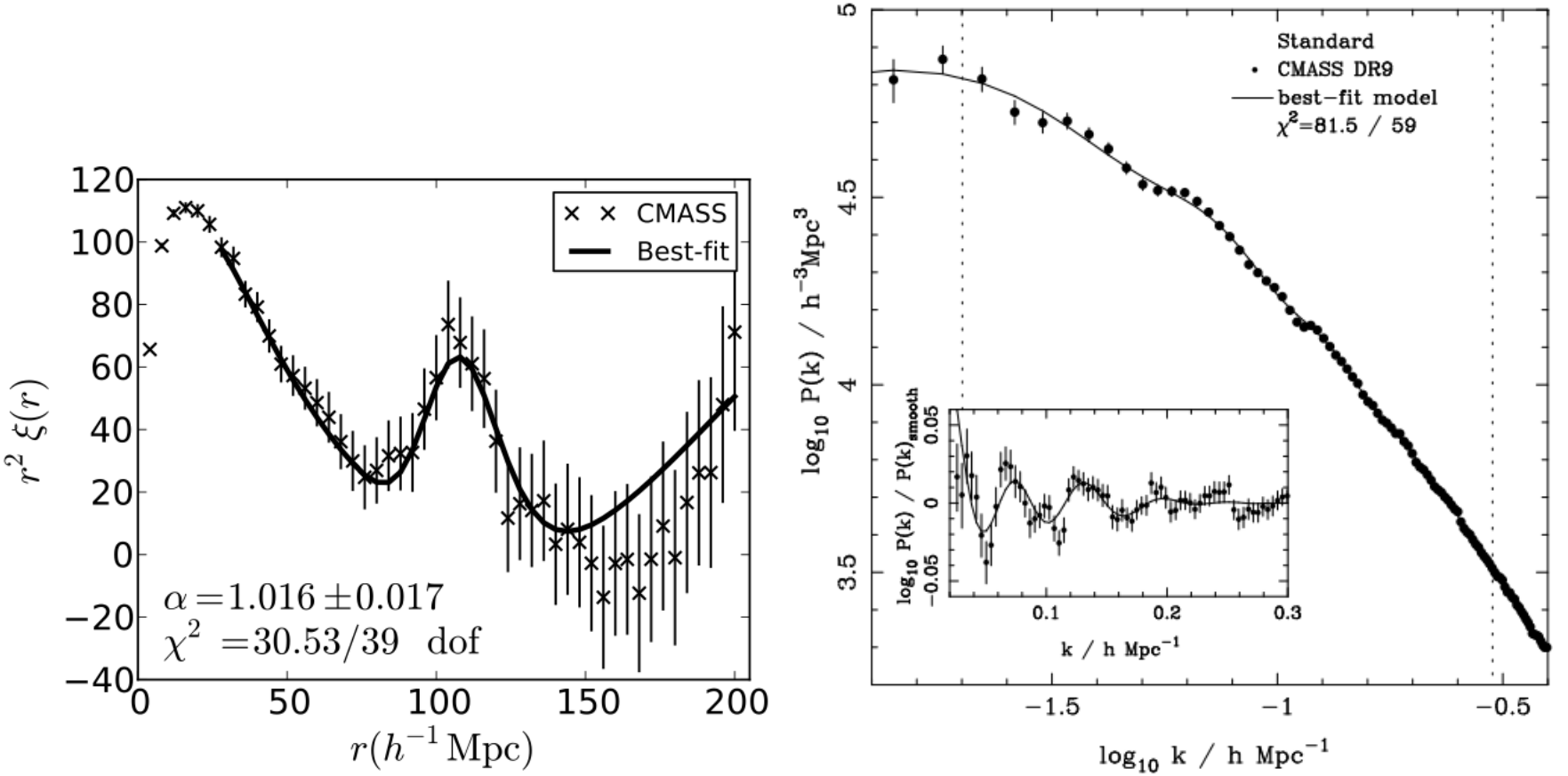


Figure 1.6: The galaxy two-point correlation function (left) and power spectrum (right) from the SDSS-III Baryon Oscillation Spectroscopic Survey Data Release 9, showing the distinct imprints of the BAO left in the clustering of galaxies. The BAO shows up as a peak in the two-point correlation function and oscillations in the power spectrum. The latter of these has been isolated in the panel of the right plot by dividing the measured power spectrum by a smooth model without the BAO features. These plots are taken from Anderson et al. (2014a) and their best-fitting model and recovered parameters are also shown.

Hubble parameter

$$H(z) = \frac{c\Delta z}{r_d}. \tag{1.106}$$

If the spherically averaged two point statistics are used, such that there is no distinction between the clustering along and perpendicular to the line of sight, the BAO peak position relates to an isostropic combination of $D_A(z)$ and $H(z)$

$$D_V(z) = \left[(1+z)^2 D_A^2(z)\frac{cz}{H(z)}\right]^{1/3}. \tag{1.107}$$

In practice, the scale of the BAO feature is defined relative to a fiducial cosmology, with corresponding angular diameter distance $D_{A,fid}(z)$, Hubble parameter $H_{fid}(z)$ and sound horizon $r_{d,fid}$. If this fiducial cosmology is used to convert the angular and redshift coordinates of a galaxy into a cartesian basis, then the measured two-point clustering of the galaxies is 'dilated' by a factor

$$\alpha = \frac{D_V(z) r_{d,fid}}{D_{V,fid}(z) r_d} \tag{1.108}$$

compared to the true clustering. If the fiducial cosmology is identical to the true cosmology then $\alpha = 1.0$. Values of $\alpha < 1.0$ and $\alpha > 1.0$ mean that features in the two point clustering are shifted to larger/smaller scales respectively. Hence the peak of the sphereically averaged BAO feature can be used to constrain cosmology via the parameter α.

In a similar vein, different scaling parameters can be defined for the clustering along and transverse to the LOS,

$$\alpha_{||} = \frac{H(z) r_d}{H_{fid}(z) r_{d,fid}}, \qquad \alpha_{\perp} = \frac{D_A(z) r_{d,fid}}{D_{A,fid}(z) r_d}. \tag{1.109}$$

Hence measurements of the clustering in these distinct directions allows the degeneracy between $D_A(z)$ and $H(z)$ present in the spherically averaged measurement to be broken.

Figure 1.7 shows a compilation of the spherically averaged distance measurements made using the BAO feature from a range of surveys, listed in Table 1.1. These measurements show excellent support for the consensus cosmological model given in Section 1.1.4, which is plotted alongside.

1.4.4 Redshift Space Distortions

Determining the 3D distribution of structure in the universe requires measurements of the angular coordinates of a galaxy on the sky, and a measure of its comoving distance from the observer. For a given galaxy this latter component cannot be measured directly, only inferred from the redshift using Eq 1.26. However, Hubble expansion is not the only contributor to the redshift measured from a galaxy, with additional components arising

Table 1.1: A compilation of high precision measurements of the spherically-averaged distance D_V/r_d measured from a variety of different surveys over a range of redshifts. In the case of multiple measurements using the same dataset at similar redshifts, the most constraining measurement is quoted. Measurements in bold indicate the optimal combination of sufficiently independent measurements that can be combined for joint likelihood analysis. From Figure 1.7 one can see that it is preferential to choose the SDSS-III BOSS DR11 measurements over those of the WiggleZ survey as the extremely tight constraints dominate the resultant joint likelihood. The method used to determine the $z = 0.15$ measurement using the SDSS-II DR7 MGS sample will be the focus of Chapters 3 and 4

z	D_V/r_d	Dataset	Reference
0.106	$\mathbf{3.060 \pm 0.120}$	**6dFGRS**	**Beutler et al. (2011)**
0.150	$\mathbf{4.470 \pm 0.160}$	**SDSS-II DR7 MGS**	**Ross et al. (2015)**
0.200	5.389 ± 0.173	SDSS-II DR7 LRG	Percival et al. (2010)
0.320	$\mathbf{8.466 \pm 0.166}$	**SDSS-III BOSS DR11**	**Tojeiro et al. (2014)**
0.350	9.106 ± 0.174	SDSS-II DR7 LRG	Padmanabhan et al. (2012)
0.440	11.546 ± 0.558	WiggleZ	Kazin et al. (2014)
0.570	$\mathbf{13.773 \pm 0.134}$	**SDSS-III BOSS DR11**	**Anderson et al. (2014b)**
0.600	14.944 ± 0.680	WiggleZ	Kazin et al. (2014)
0.730	16.929 ± 0.579	WiggleZ	Kazin et al. (2014)

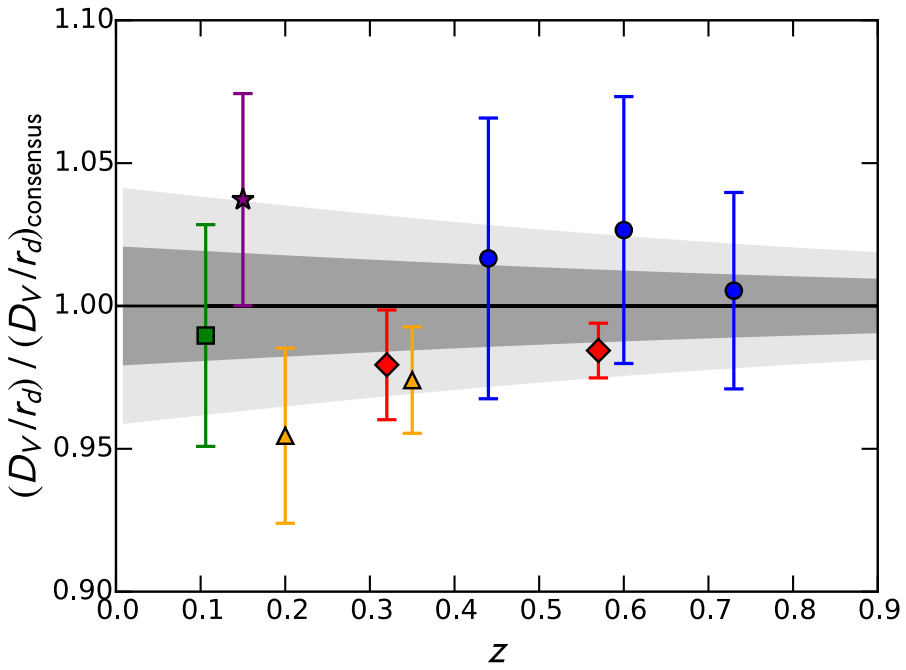


Figure 1.7: The BAO distance ladder as a function of redshift compared to that predicted from the consensus cosmology of Section 1.1.4, measured from a variety of surveys listed in Table 1.1. Green, purple, orange, red and blue points correspond to measurements from the 6dFGRS, SDSS-II DR7 MGS, SDSS-II DR7 LRG, SDSS-III BOSS DR11 and WiggleZ surveys. The grey and light grey regions denote the 1 and 2σ constraints based on the errors on the consensus cosmology.

mainly from the peculiar velocities of galaxies as they infall into potential wells. If the peculiar velocity of a galaxy has components along the line-of-sight with respect to an observer, this then means that a galaxy position calculated from its redshift will be different from its true position.

In terms of the clustering of galaxies, this manifests as a squashing of otherwise spherical overdensities; galaxies between an overdensity and the observer will be falling towards the potential and hence have a peculiar velocity that acts to make them appear further from the observer than their true distance. Similarly galaxies beyond the overdensity will seem closer. The net effect of this is a squashing of the overdensity along the line-of-sight. In the same way, an underdense region will appear elongated along the line of sight. This effect is known as Redshift Space Distortions (RSD)

In an otherwise spherical distribution, RSD creates an asymmetry between the radial and transverse directions that in turn induces higher order moments in the clustering of galaxies. In the absence of RSD both the two-point correlation function and power spectrum are spherically symmetric, whilst RSD creates dipolar, hexadecapolar etc. moments.

Effect on the Density Field

Kaiser (1987) was the first to express the effects of RSD on the two-point clustering of matter. The cornerstone of the Kaiser model is that the number density of sources in a volume element is conserved when mapping from a location $\boldsymbol{r}$ in real-space to $\boldsymbol{s}$ redshift-space, such that

$$\rho(\boldsymbol{x})d^3r = \rho_s(\boldsymbol{x}_s)d^3s \tag{1.110}$$

where ρ and ρ_s are the number densities in real and redshift space, whilst $d^3\boldsymbol{r}$ and $d^3\boldsymbol{s}$ are the real and redshift space volume elements. The Jacobian of the transformation between real and redshift coordinates is then

$$J \equiv \left|\frac{d^3r}{d^3s}\right| = \frac{dr}{ds}\frac{r^2}{s^2}. \tag{1.111}$$

The latter expression follows from expressing the volume elements as infinitesimally thin shells, which have the same angular size in real and redshift space. Neglecting subdominant terms such as gravitational redshift, the redshift-space distance of a galaxy can be written in terms of the peculiar velocity $\boldsymbol{v}$ (in units of the Hubble parameter) and real space distance as,

$$s = r + \boldsymbol{v}\cdot\hat{\boldsymbol{r}}, \tag{1.112}$$

and hence the Jacobian of the coordinate transformation from real to redshift space is

$$J = \left(1 + \frac{\partial}{\partial r}[\boldsymbol{v}\cdot\hat{\boldsymbol{r}}]\right)^{-1}\left(1 + \frac{\boldsymbol{v}\cdot\hat{\boldsymbol{r}}}{r}\right)^{-2}. \tag{1.113}$$

To proceed, Kaiser (1987) made an important assumption, that the first term of the Jacobian dominates over the second term. The argument for this stems from the idea that the survey encloses the source of the peculiar motion. In Fourier space, the first, derivative term $\partial/\partial r[\boldsymbol{v}\cdot\hat{\boldsymbol{r}}] \rightarrow k[\boldsymbol{v}\cdot\hat{\boldsymbol{r}}]$, whilst the second term depends on $[\boldsymbol{v}\cdot\hat{\boldsymbol{r}}]/r$. Hence if the wavelength of the perturbations of interest satisfy $kr >> 1$, which is true if the source of the peculiar motions is within the survey, the second term can be neglected and the Jacobian Taylor expanded about $\boldsymbol{v} = 0$.

$$J \approx \left(1 - \frac{\partial}{\partial r}[\boldsymbol{v}\cdot\hat{\boldsymbol{r}}]\right) \tag{1.114}$$

Substituting the Jacobian into Eq. 1.110, and writing the number densities in terms of overdensities gives

$$\delta_s = [1+\delta]\left(1 - \frac{\partial}{\partial r}[\boldsymbol{v}\cdot\hat{\boldsymbol{r}}]\right) - 1 \approx \delta - \frac{\partial}{\partial r}[\boldsymbol{v}\cdot\hat{\boldsymbol{r}}]. \tag{1.115}$$

The last expression is derived from the first by expanding to linear order (neglecting the term containing both δ and $\partial/\partial r[\boldsymbol{v}\cdot\hat{\boldsymbol{r}}]$ as these are both much smaller than 1) and allows direct correspondence between the real and redshift space overdensities based on the peculiar velocity.

Effect of RSD on the Two-point Clustering

In the linear regime, the redshift space power spectrum is relatively simple to relate to the real space power spectrum. Taking the Fourier transform of Eq. 1.115 and substituting in the peculiar velocity from Eq 1.82, keeping in mind the convention used in this chapter to write the peculiar velocity in units of the Hubble parameter (which cancels out the aH factor in Eq. 1.82),

$$\delta_s(\boldsymbol{k}) = \delta(\boldsymbol{k}) - if\int d^3r e^{-i\boldsymbol{k}\cdot\boldsymbol{r}}\frac{\partial}{\partial r}\left[\int \frac{d^3\boldsymbol{k'}}{(2\pi^3)}e^{i\boldsymbol{k'}\cdot\boldsymbol{r}}\delta(\boldsymbol{k'})\hat{\boldsymbol{k'}}\cdot\hat{\boldsymbol{r}}\right]. \tag{1.116}$$

Under the distant-observer approximation, the vector product $\boldsymbol{v}\cdot\hat{\boldsymbol{r}} \leftarrow \boldsymbol{v}\cdot\hat{r}$, that is, the direction vector consists of only a radial component. Using this approximation, and a little maths,

$$\delta_s(\boldsymbol{k}) = \delta(\boldsymbol{k})[1 + f\mu_{\boldsymbol{k}}^2], \tag{1.117}$$

where $\mu_k = \hat{\boldsymbol{k}}\cdot\hat{r}$.

As shown previously, the galaxy overdensity is a biased version of the matter overdensity, which is the dominant source of the peculiar velocities. As such the quantity of interest is actually

$$\delta_s(\boldsymbol{k}) = \delta(\boldsymbol{k})[1 + \beta\mu_{\boldsymbol{k}}^2], \tag{1.118}$$

where $\beta = f/b$. In turn, the redshift space galaxy power spectrum is

$$P_g^s(\boldsymbol{k}) = [1 + \beta\mu_{\boldsymbol{k}}^2]^2 P_g(\boldsymbol{k}). \tag{1.119}$$

The dependence on $\mu_{\boldsymbol{k}}$ is what induces higher order multipoles in the measured two-point galaxy clustering. It is common to expand the power spectrum and two-point correlation function in terms of these multipoles using the Legendre polynomials, $\mathcal{P}_\ell(\mu)$, such that

$$P(k, \mu_k) = \sum_{l=0}^{\infty} P_\ell(k)\mathcal{P}_\ell(\mu_k) \text{ and} \tag{1.120}$$

$$\xi(r, \mu_r) = \sum_{l=0}^{\infty} \xi_\ell(r)\mathcal{P}_\ell(\mu_r). \tag{1.121}$$

For the power spectrum the individual multipoles can be expressed, in linear theory, as functions of the real space power spectrum via

$$P_g^{0,s}(k) = \int_0^1 d\mu_k [1 + \beta\mu_{\boldsymbol{k}}^2]^2 P_g(\boldsymbol{k}) = \left[1 + \frac{2}{3}\beta + \frac{1}{5}\beta^2\right] P_g(k), \tag{1.122}$$

$$P_g^{2,s}(k) = 5\int_0^1 d\mu_k \mathcal{P}_2(\mu_k)[1 + \beta\mu_{\boldsymbol{k}}^2]^2 P_g(\boldsymbol{k}) = \left[\frac{4}{3}\beta + \frac{4}{7}\beta^2\right] P_g(k), \tag{1.123}$$

$$P_g^{4,s}(k) = 9\int_0^1 d\mu_k \mathcal{P}_4(\mu_k)[1 + \beta\mu_{\boldsymbol{k}}^2]^2 P_g(\boldsymbol{k}) = \frac{8}{35}\beta^2 P_g(k). \tag{1.124}$$

Hamilton (1992) derived the similar expressions for the redshift space, two-point correlation function multipoles. The μ dependence of the redshift space clustering allows one to measure the value of f, and hence test General Relativity, if they measure the clustering at different orientations with respect to the line-of-sight, i.e., the full anisotropic clustering as opposed to the spherically averaged clustering. The most common method of determining f is to use the multipoles. In the linear regime this is as simple as determining the ratio of the first and second multipoles, however this does highlight an important effect. In the linear regime the growth rate is completely degenerate with the galaxy bias. This degeneracy can be partially broken by looking at non-linear scales where the redshift space clustering depends on more than just β, however a strong degeneracy still remains, which creates difficulties in constraining f. An additional degeneracy arises from the amplitude of the dark matter power spectrum itself which on large scales acts in a similar manner to galaxy bias. As such, with current data and RSD models, it is almost impossible to disentangle the inherent amplitude of the dark matter power spectrum, parameterised by σ_8, from b and f. It is common practice to absorb σ_8 into the constraints, such that modern RSD analyses measure the combinations $b\sigma_8$ and $f\sigma_8$ (Percival & White, 2009).

Overall, the above derivation has a somewhat serious flaw. It only holds on large scales, where the linear versions of the real to redshift space transformation and continuity equation hold. However, in a somewhat cruel turn of events, it is observationally

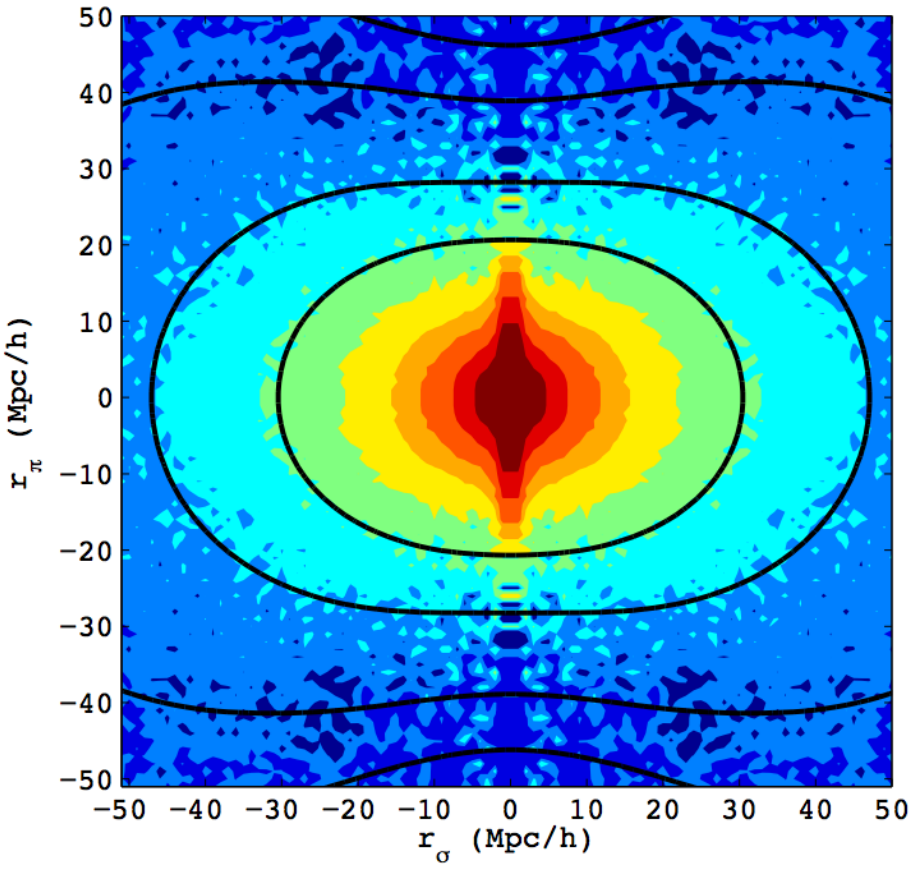


Figure 1.8: The two-point correlation function along and perpendicular to the line-of-sight as measured by Reid et al. (2012) using SDSS-III BOSS DR9 data. Also shown is their best-fitting RSD model. The figure clearly shows the anisotropy in the clustering induced by the bulk motions of galaxies towards overdensities, and the non-linear Finger-of-God arising from the peculiar motions of galaxies within virialised dark matter halos.

easiest to measure the effects of peculiar velocities on small scales, where the effects of Hubble expansion are subdominant and the signal is corresponding higher. This means that detailed non-linear models of the redshift space clustering are required to make the most of the clustering measurements obtained from modern surveys. An example of such a model is presented in Chapter 4.

In the last decade much attention has been given to measuring the RSD signal in the galaxy clustering from a variety of LSS surveys (Percival et al., 2004; Guzzo et al., 2008; Beutler et al., 2012; Samushia et al., 2012; Oka et al., 2014; Beutler et al., 2013; Chuang et al., 2013; Samushia et al., 2014; Sánchez et al., 2014; Blake et al., 2011a,b). A compilation of these measurements alongside the predictions from General Relativity is shown in Chapter 4, where excellent agreement can be seen between the two. Figure 1.8 shows the galaxy correlation function along and perpendicular to the line-of-sight using

SDSS-III BOSS DR9 data from Reid et al. (2012) along with their best-fit model. In this figure the effects of RSD can be clearly seen. The clustering is squashed along the line-of-sight as detailed previously, creating the higher order multipoles. An additional feature arising from RSD, that has not been covered in the linear derivation above can be seen on very small scales. These are the 'Fingers-of-God', non-linear RSD effects arising from the peculiar motions of galaxies within virialised halos. These galaxies have already fallen into the potential wells and are moving with respect to the common centre of mass of the bound structure. The motions of these galaxies creates an elongation of the clustering along the line-of-sight.

1.5 Summary and Thesis Outline

This introductory chapter has provided an overview of the current 'state-of-play', of cosmology, presenting and deriving the concordance model of cosmology and showing ways in which the large scale structure of the universe can be used to probe cosmology. Whilst not fully comprehensive, the topics covered in this chapter serve to put the remainder of this thesis in context.

Within this introduction the need for accurate measurements of the power spectrum and its covariance matrix has been identified, and fast, accurate methods of producing simulations is a step towards making more precise measurements of large scale structure in the future. Chapter 2 will contain an overview of previous methods to create dark matter fields based on solutions to the equations for the evolution of CDM perturbations,. This will set the scene for L-PICOLA, a new method for producing fast simulations of the large scale dark matter distribution of the universe. This code is thoroughly demonstrated and tested for both accuracy and speed within this chapter.

Chapters 3 and 4 will present the methods used to produce a new set of BAO and RSD measurements at low redshift for cosmological constraints. As seen in Figure 1.7, the consensus cosmology detailed in this introduction has greatest uncertainty at low redshift and hence it is at low redshift that LSS measurements have the largest relative constraining power. Chapter 3 will show and justify the choice of data used to make the measurements before showing how L-PICOLA was applied to this dataset to create a set of accurate mock catalogues. Large ensembles of mock catalogues are the primary method of estimating the covariance matrix and testing the fitting methodology of large scale structure measurements. Chapter 3 presents these tests and the mocks which ultimately enable cosmology constraints to be obtained from the data. Fitting the BAO and RSD signals in the data and the subsequent constraints form the basis of Chapter 4 where the results are placed in the context of other BAO and RSD measurements and used to improve current measurements of the nature of dark energy.

Chapter 5 presents a novel, optimal, method for evaluating the covariance matrix without the need to run ensembles of simulations that cover the full survey volume. As future surveys cover larger and larger cosmological volumes, it becomes unfeasible to simulate the full survey, with a resolution down to the required halo mass limit, enough times to generate accurate estimates of the covariance matrix. Instead, Chapter 5 shows how an optimum way to estimate the covariance matrix is to combine small-volume simulations that capture the non-linear information, with large-scale analytic predictions of the covariance matrix, taking into account the effects of modes larger than the small-volume simulations. This will enable future LSS surveys to reach their full potential. In complement to the other chapters in this work, and to round out the thesis, this uses the simulations from Chapter 3, created via the code in Chapter 2, previously developed to obtain the BAO and RSD constraints in Chapter 4.

Finally, Chapter 6 gives an overview of the thesis and concludes the work here. Possible future directions for this work will also be given here.

Chapter 2

L-PICOLA: A New Code for Fast Dark Matter Simulation

Analysis of large scale structure allows one to probe the universe over a wide redshift range, measuring its expansion history and providing the most robust route to measuring its late-time evolution. Over the last decade, large sky-area galaxy surveys such as the Sloan Digital Sky Survey (SDSS; York et al. 2000; Eisenstein et al. 2011), 2dF Galaxy Redshift Survey (2dFGRS; Colless et al. 2001, 2003), 6dF Galaxy Redshift Survey (6dFGRS, Jones et al. 2004, 2009) and WiggleZ survey (Drinkwater et al., 2010) have led to robust measurements of large scale structure and provided a wealth of cosmological information.

In particular, as shown in Chapter 1, measurements of the BAO scale (Cole et al., 2005; Eisenstein et al., 2005) provide us with a standard ruler, allowing us to measure the accelerated expansion of the universe, whilst RSD (Kaiser, 1987) provide a direct probe of the growth of structure and the fidelity of General Relativity. These probes have become more and more accurate in recent years, with Anderson et al. (2014b) providing a 1% measurement of the distance scale to $z = 0.57$, the most precise from a galaxy survey to date. However, these measurements and their errors require intimate knowledge of the statistical and systematic distributions from which they are drawn. This need will only be exacerbated as future surveys, such as the Large Sky Synoptic Telescope (LSST; Ivezic et al. 2008) and Euclid survey (Laureijs et al., 2011), strive for even greater precision.

The most accurate method of estimating the statistical errors within an LSS survey is to use large numbers of realistic mock galaxy catalogues that mimic the observed distribution of galaxies within the survey volume. Ideally these simulations would take the form of fully non-linear N-Body simulations, with accurate small scale clustering, covering the whole volume of the galaxy survey. However, for current surveys, recent studies (Dodelson & Schneider, 2013; Taylor et al., 2013; Percival et al., 2014) show that $\mathcal{O}(1000)$

mocks are required to obtain an accurate numerical estimate of the covariance matrix with sub-dominant errors compared to the statistical errors themselves. Higher precision measurements in the future may require many more. Instead there have been many studies looking at fast methods of producing simulations that enable one to produce mocks hundreds of times faster than an Tree-PM N-Body simulation, at the cost of reduced small scale clustering accuracy.

To this end, this chapter presents a fast, distributed-memory, planar-parallel code, L-PICOLA, which can be used to generate and evolve a set of initial conditions into a dark matter field much faster than a full non-linear N-Body simulation. The code has been created with emphasis on maximising speed, memory conservation and ease of use. Additionally, L-PICOLA has the ability to include primordial non-Gaussianity in the simulation and simulate the past lightcone at run-time, with optional replication of the simulation volume. The accuracy, speed and scalability of this code, alongside the additional features that have been implemented, make it extremely useful for both current and next generation large-scale structure surveys. In fact, as will be detailed in Chapters 3 and 4, this code has already been used to produce measurements of the BAO and RSD signals from a subset of luminous red galaxies drawn from the Sloan Digital Sky Survey Data Release 7 Main Galaxy Sample.

This chapter is structured as follows: Section 2.1 provides a small review of some current methods for simulating dark matter fields, with particular emphasis on the techniques used within L-PICOLA. Section 2.2 introduces the code itself, while Section 2.3 details the steps taken to parallelise the code for a distributed-memory machine. Sections 2.4 and 2.5 detail the additional features that have been included in L-PICOLA, beyond the ability to create standard snapshot simulations. In particular, Section 2.5 validates the need for lightcone simulations and performs a rigorous test of the implementation within the code. Section 2.6 compares the accuracy of L-PICOLA to a single-step 2nd order Langrangian perturbation theory (2LPT) simulation and a full N-Body simulation. The effects on the clustering accuracy of several of the free parameters that are used to speed up the convergence of the code are also tested in this section. In Section 2.7 the speed of L-PICOLA is compared with 2LPT simulations, and the scaling of different segments of L-PICOLA itself is presented, whilst Section 2.8 gives a brief overview of the memory requirements of L-PICOLA. Finally Section 2.9 concludes this chapter and discusses further improvements that can be made to the code.

Unless otherwise stated, a fiducial cosmology given by $\Omega_m = 0.317$, $\Omega_b = 0.049$, $h = 0.67$, $\sigma_8 = 0.83$, and $n_s = 1.0$ is assumed. Also, unless otherwise stated, all simulations presented use a number of mesh cells equal to the number of particles, the COLA method with modified COLA timestepping, $nLPT = -2.5$ and 10 linearly

spaced timesteps. These L-PICOLA-specific parameters are stated here for completeness but are explained within this chapter.

2.1 Simulating Late-Time Cold Dark Matter

The first port-of-call in the production of a mock galaxy catalogue is the simulation of a dark matter field at the redshift of the observations. This involves taking the state of the dark matter field shortly after recombination, usually done using the matter power spectrum, and evolving it to the required time. There is a huge variety of methods for doing this, which differ greatly in their complexity, accuracy and speed. It is often the case that accuracy in individual simulations will be sacrificed in order to obtain sufficient numbers to estimate the covariance matrix. As such there is perhaps no correct method, though some are certainly more suitable than others. This section will provide an overview of the current methods for simulating dark matter used in producing mock catalogues.

2.1.1 Gaussian/Lognormal Realisations

Historically, a common method of producing a dark matter distribution based on some input matter power spectrum is through Gaussian or Lognormal realisations of the density field.

Gaussian realisations of the density field have been widely studied in the literature (Peacock & Heavens, 1985; Bardeen et al., 1986; Jensen & Szalay, 1986; Couchman, 1987; Lumsden et al., 1989). As covered in Chapter 1 it is expected that the primordial density fluctutations arising from inflation are nearly perfectly Gaussian. As such, a good first approximation of the late-time density field is that it is simply a Gaussian realisation, with independent real and imaginary parts drawn from a distribution with variance given by the input power spectrum., i.e,

$$\mathcal{P}(\delta_{\boldsymbol{k}}) = \frac{1}{\sqrt{2\pi^2 P_{\boldsymbol{k}}}} e^{-\frac{\delta_{\boldsymbol{k}}^2}{2P_{\boldsymbol{k}}}}. \tag{2.1}$$

This results in a density field with random phases and an amplitude drawn from a Rayleigh distribution.

The simple formalism of the Gaussian realisation allows for analytic solutions to be obtained for many of the properties of the dark matter density fields, such as the number density of density peaks of a given height. They are also extremely quick to produce, requiring one to simply draw from the above distribution within some k-space grid.

Unfortunately, the approximation that the late-time density field is Gaussian is only true for the largest scales. In even mildly non-linear regimes the effect of gravitational collapse introduces non-Gaussianity into the density distribution (Fry, 1986). This results

in a skewed density distribution caused by a relatively small number of extremely high peaks in the density field. Subsequently, Coles & Barrow (1987) and Coles & Jones (1991) introduced a transformation of the Gaussian field into a lognormal field, which better describes the density probability distribution function. In this Lognormal method

$$\mathcal{P}(\delta_{\boldsymbol{k}}) = \frac{1}{\sqrt{2\pi^2 P_{\boldsymbol{k}}}} e^{-\frac{(\log\delta_{\boldsymbol{k}})^2}{2P_{\boldsymbol{k}}}} \frac{1}{\delta_{\boldsymbol{k}}}. \tag{2.2}$$

Much like the Gaussian realizations, this method is also easy to model analytically, and fast to implement.

Both Gaussian and Lognormal realisations have been used to produce mock catalogues for analysis of large scale structure data from the 2DF galaxy redshift survey by Percival et al. (2001) and Cole et al. (2005). However, with the advent of modern computing power, these methods have been far surpassed by more accurate methods that, whilst slower to compute and more analytically complex, are still relatively easy to implement and whose parallelised implementations can produce large ensembles of mocks very quickly. Gaussian realizations generally remain in use today to generate the initial conditions for more detailed N-Body simulations. For instance, a common method of producing a set of initial conditions, and one that is adopted in L-PICOLA, is to draw Gaussian density perturbations from some input power spectrum, then use analytic approaches such as Lagrangian perturbation theory to slightly perturb the particle positions at high redshift, before iteratively evolving the density field. Both Lagrangian perturbation theory and iterative methods for evolving the density field will be presented in the next section.

2.1.2 Lagrangian Perturbation Theory

Using the notation of Scoccimarro (1998) (see also Moutarde et al. 1991 and Bouchet et al. 1995), cold dark matter particles evolving over cosmological time in an expanding universe follow the equation of motion (EOM)

$$\frac{d^2\boldsymbol{\Psi}}{d\tau^2} + \mathcal{H}(\tau)\frac{d\boldsymbol{\Psi}}{d\tau} + \nabla\Phi = 0, \tag{2.3}$$

where Φ is the gravitational potential, $\mathcal{H}(\tau) = \frac{dlna}{d\tau}$ is the conformal Hubble parameter and a is the scale factor. $\boldsymbol{\Psi}$ is the displacement vector of the particle. This new variable has been introduced compared to Eq. 1.69 to account for the discretisation of the density field. The EOM for the CDM particles can be solved either iteratively, or if some approximations are used, in a single-step. One such single step solution is Lagrangian Perturbation Theory

Due to its speed and ease of implementation (at least to second order) Lagrangian perturbation theory (LPT) has proven to be a very popular method for solving the dark

matter equation of motion. It involves a perturbative expansion of the displacement vector $\mathbf{\Psi}$ of each particle. In LPT, the physical meaning of the displacement vector $\mathbf{\Psi}$, is that it relates a particle's Eulerian position $\boldsymbol{x}(\tau)$ to its initial, Lagrangian position, $\boldsymbol{q}$, via

$$\boldsymbol{x}(\tau) = \boldsymbol{q} + \Psi(\boldsymbol{q}, \tau). \tag{2.4}$$

By taking the divergence of the equation of motion and using the Poisson equation,

$$\nabla_x \cdot \left(\frac{d^2\mathbf{\Psi}}{d\tau^2} + \mathcal{H}(\tau)\frac{d\mathbf{\Psi}}{d\tau} \right) = -\frac{3}{2}\Omega_{m,0}\mathcal{H}(\tau)\delta(\tau). \tag{2.5}$$

Here $\Omega_{m,0}$ is the matter density at $\tau = 0$, whilst $\delta(\tau)$ is the local overdensity. Lagrangian perturbation theory seeks to solve this equation by perturbatively expanding the displacement vector,

$$\mathbf{\Psi} = \mathbf{\Psi}^{(1)} + \mathbf{\Psi}^{(2)} + \ldots, \tag{2.6}$$

If the continuity equation is applied, $\rho(\boldsymbol{x}, t)d^3x = \rho(\boldsymbol{q}, 0)d^3q$, which states that a mass element d^3q centred at $\boldsymbol{q}$ at time zero becomes a mass element d^3x, centred at $\boldsymbol{x}$, at time t, then to first order

$$\nabla_q \cdot \mathbf{\Psi}^{(1)} = -D_1(\tau)\delta_L(\boldsymbol{q}). \tag{2.7}$$

This is the well known Zel'dovich appoximation (ZA; Zel'dovich 1970). $D_1(\tau)$ is the linear growth factor which was introduced in Chapter 1, $\delta_L(q)$ is the linear overdensity field and the divergence has been rewritten as a function of $\boldsymbol{q}$ by using the fact that they are related via the Jacobian of the transformation from $\boldsymbol{x}$ to $\boldsymbol{q}$, i.e., $\nabla x_i = (\delta_{ij} + \partial\mathbf{\Psi}_i/\partial\boldsymbol{q}_j)^{-1}\nabla q_j$. Solving to second order introduces corrections to the first order displacement of the form

$$\nabla_q \cdot \mathbf{\Psi}^{(2)} = \frac{1}{2}D_2(\tau)\sum_{i\neq j}\left(\mathbf{\Psi}^{(1)}_{i,i}\mathbf{\Psi}^{(1)}_{j,j} - \mathbf{\Psi}^{(1)}_{i,j}\mathbf{\Psi}^{(1)}_{j,i} \right), \tag{2.8}$$

where, for brevity, the definition $\mathbf{\Psi}_{i,j} = \partial\mathbf{\Psi}_i/\partial\boldsymbol{q}_j$ has been used. Bouchet et al. (1995) provide a good approximation for $D_2(\tau)$, the second order growth factor, for a flat universe with non-zero cosmological constant

$$D_2(\tau) \approx -\frac{3}{7}D_1^2(\tau)\Omega_m(\tau)^{-1/143}. \tag{2.9}$$

For further computational ease, one can define two Lagrangian potentials, $\mathbf{\Psi}^{(i)} = \nabla_q\phi^{(i)}$, such that Eq. 2.4 becomes

$$\boldsymbol{x}(\tau) = \boldsymbol{q} - D_1\nabla_q\phi^{(1)} + D_2\nabla\phi^{(2)}, \tag{2.10}$$

and the Lagrangian potentials are obtained by solving the corresponding pair of Poisson equations derived from Eq. 2.7 and Eq. 2.8 respectively,

$$\nabla_q^2\phi^{(1)} = \delta_L(\boldsymbol{q}). \tag{2.11}$$

$$\nabla_q^2 \phi^{(2)} = \sum_{i>j} \left(\phi_{,i}^{(1)} \phi_{,j}^{(1)} - (\phi_{,j}^{(1)})^2 \right). \tag{2.12}$$

The strength of Lagrangian perturbation theory is that it provides an exact solution in the large scale limit and as long as the density contrast is small enough for the perturbative expansion of the displacement to remain valid. However this method fails to fully capture the non-linear evolution of the dark matter particles, despite becoming much more complex after second-order, and breaks down at 'shell-crossing', where two particle trajectories intersect. Even so it has still seen usage in recent times as a method for producing mock catalogues for the BOSS survey, with a small correction to limit shell-crossing (Scoccimarro, 1998; Manera et al., 2013, 2015).

As the LPT solution to second-order is more accurate on linear and quasi-linear scales, and hence at high redshift, it is often used to generate high-redshift initial conditions for use in more non-linear, iterative N-Body codes. The COLA method decribed later, and the code that is the main focus of this chapter, L-PICOLA, use the exact large scale solution provided by second-order LPT to speed up the iterative solution to the dark matter EOM.

Other methods exist that make use of Lagrangian perturbation theory but extend its validity further into the non-linear regime for use in producing fast simulations of dark matter. Augmented Lagrangian Perturbation Theory (Kitaura & Heß, 2013; Heß et al., 2013) and its implementation, PATCHY (Kitaura et al., 2014), combine the large scale analytic 2LPT solutions with the small scale sphereical collapse model to evolve dark matter particles, connecting the two using a Gaussian smoothing on quasi linear scales. Similarly, the 'EZmocks' method (Chuang et al., 2015) simulates a dark matter field using the ZA before matching the probabilty distribution of the simulations to a non-linear, iterative N-Body simulation to improve the non-linear clustering. Finally, the PINOCCHIO code (Monaco et al., 2002, 2013), uses the ZA and the phenomenon of shell-crossing to simulate and identify the formation of dark matter halos. All of these methods improve on the accuracy of standard LPT solutions with minimal computational cost, although they still lack small scale accuracy compared to the non-linear solution provided by solving the dark matter EOM iteratively.

2.1.3 N-Body Methods

Perhaps the most ubiquitous method of simulating dark matter is through the use of N-Body techniques, which convert the continuous dark matter density distribution into a set of discrete 'particles' each with some mass, position and velocity. The evolution of the dark matter density perturbations can then be simulated by computing the position and velocity of each of the particles as the universe evolves. N-Body particles evolving under

the force of gravity follow an equation of motion based on the second order equation for the evolution of a dark matter density perturbation previously seen in Eq. 2.3

The equation of motion can be further discretised in time. The most common example of this is the Kick-Drift-Kick/Leapfrog method (Quinn et al., 1997). In this method, the dark matter EOM is solved in a series of timesteps, and at each iteration the particle velocities and positions are updated based on the gravitational potential felt by each particle. Particle velocities are calculated from the displacements and updated to the nearest half-integer timestep. The particle positions are then updated to the nearest integer timestep using the previous velocity. In this way the particle velocities and positions are never (except at the beginning and end) calculated for the same point in time but rather 'leapfrog' over each other with the next iteration of the velocity dependent on the position from the previous iteration and so on.

Mathematically, this can be described via.

$$\boldsymbol{v}_{i+1/2} = \boldsymbol{v}_{i-1/2} - \nabla\Phi_i \Delta a_1, \tag{2.13}$$

$$\boldsymbol{r}_{i+1} = \boldsymbol{r}_i + \boldsymbol{v}_{i+1/2}\Delta a_2. \tag{2.14}$$

Δa_i encapsulates the time interval and appropriate numerical factors required to convert the displacement to a velocity and the velocity to a position. Quinn et al. (1997) evaluate these as

$$\begin{aligned} \Delta a_1 &= \frac{H_0}{a_i} \int_{a_{i-1/2}}^{a_{i+1/2}} \frac{da}{a^2 H(a)}, \\ \Delta a_2 &= H_0 \int_{a_i}^{a_{i+1}} \frac{da}{a^3 H(a)}. \end{aligned} \tag{2.15}$$

When solving the dark matter equation of motion iteratively, a method of evaluating the gravitational potential felt by each particle at each timestep is still required. There are roughly three ways of doing this: directly summing the contributions from each particle in the simulation; evaluating the density distribution and the gravitational potential on a grid, or some combination of the two.

The first of these is trivial to implement and exactly recovers the gravitational potential, however quickly becomes unfeasible for large numbers of particles as it scales with the number of particles as $\mathcal{O}(N^2)$. The particle-mesh method on the other hand assigns particles to a mesh and solves the force via a Fourier transformation. Hence this reduces the scaling to $\mathcal{O}(N_m \log N_m)$ where N_m is the number of mesh cells used. The downside to this is a loss of accuracy as neither the density nor the potential are known at the exact location of each particle, only at the centre of the mesh cell. The value at each particle's location must be interpolated. This means that once again accuracy is being sacrificed for speed.

This can be overcome somewhat by using a suitably large number of mesh cells, and its speed is such that it still sees much use today in producing fast simulations of dark matter fields. The Quick Particle Mesh method and code (White et al., 2014) is one such example, where the particle-mesh algorithm is used with a small number of timesteps to iteratively solve the dark matter EOM. The COLA method and the implementation L-PICOLA detailed in this chapter also use the particle-mesh method to evaluate the small scale non-linear clustering, whilst the large scales are solved by LPT. The particle-mesh method will be presented in more detail in the next subsection.

The particle-mesh algorithm is fast but suffers from poor force resolution below the mesh scale, whilst the particle-particle summation provides an exact solution to the gravitational potential, but is slow. Another method is to combine these two algorithms into what is called the particle-particle particle-mesh or P3M algorithm. This algorithm is used by the well-known N-Body code GADGET-2 (Springel, 2005). The long-range potential is solved via the use of the mesh and Fourier transformations, whilst the small scale force is solved by a relatively fast summation algorithm known as the tree-algorithm. Constructing this tree uses the Barnes-Hut algorithm (Barnes & Hut, 1986), in which the particles are grouped into nodes and subnodes, such that the smallest subnodes contain only 0 or 1 particles. The summation then proceeds by summing over all distinct groups of particles, but at each nodal level approximating the potential as being at the centre of mass of the constituent subnodes. This reduces the particle-particle summation from an $\mathcal{O}(N^2)$ to an $\mathcal{O}(N \log N)$ algorithm.

Using this particle-particle summation combats the problem of the particle-mesh algorithm, but still at considerable cost to speed, usually by several orders of magnitude. This severely limits the use of fully non-linear algorithms such as those used in GADGET-2 for running ensembles of mock catalogues. These runs are often used, however, as the basis against which the faster methods are measured, and indeed the new code presented in this Chapter will be compared to GADGET-2 simulations in Section 2.6.

Particle-Mesh Algorithm

Provided here is a brief overview of the Particle-Mesh (PM) algorithm as a basis for the implementation used in L-PICOLA. The exact algorithm is based on the publicly available PMCODE found at `http://astro.nmsu.edu/~aklypin/PM/pmcode/`.

In the PM method a mesh is placed over the dark matter particles and the gravitational forces are solved at each mesh point. These are then interpolated to find the force at the position of each particle. The gravitational potential is then related to the additional velocity, and resultant displacement, for each particle. This is performed iteratively over a series of small timesteps. For N_m mesh points and N particles, this means that at each

iteration one only needs to perform N_m force calculations, which is much faster than a direct summation of the contribution to the gravitational force from each individual particle (at least for all practical applications, where $N \approx N_m$).

At each iteration the following steps are used to calculate the particle displacement:

1. Use the Cloud-in-Cell linear interpolation method to assign the particles to the mesh, thereby calculating the mass density, $\rho(\boldsymbol{x})$, at each mesh point.

2. Use a Fast Fourier Transform (FFT) to Fourier transform the density and solve the comoving Poisson equation in Fourier space.

$$k^2\phi(\boldsymbol{k}) = \frac{3}{2}\frac{\Omega_{m,0}}{a}(\rho(\boldsymbol{k}) - 1). \tag{2.16}$$

   In L-PICOLA the FFTW-3 Discrete Fourier Transform routines are used to compute the Fourier transforms. This library is freely available from `http://www.fftw.org/`.

3. Use the gravitational potential and an inverse-FFT to generate the force in each direction in real-space.

$$F(\boldsymbol{k}) = \boldsymbol{k}\phi(\boldsymbol{k}). \tag{2.17}$$

4. Calculate the acceleration each particle receives in each direction, again using the Cloud-in-Cell interpolation method to interpolate from the mesh points.

COLA

The COLA method (Tassev et al., 2013) provides a much more accurate solution to Eq. 2.3 than 2LPT, at only a moderate ($\sim 3\times$) reduction in speed. It does this by utilising the first and second-order Lagrangian displacements, which provide an accurate solution at large, quasi-linear scales, and iteratively solving for the resultant, non-linear component. By switching to a frame of reference comoving with the particles in Lagrangian space, the dark matter equation of motion can be split as follows,

$$T[\boldsymbol{\Psi}_{res}] + T[D_1]\boldsymbol{\Psi}_1 + T[D_2]\boldsymbol{\Psi}_2 + \nabla\Phi = 0, \tag{2.18}$$

where,

$$T[X] = \frac{d^2X}{d\tau^2} + \mathcal{H}\frac{dX}{d\tau}. \tag{2.19}$$

$\boldsymbol{\Psi}_{res}$ is the remaining displacement when the quasi-linear 2LPT displacements are subtracted from the full, non-linear displacement each particle should actually feel.

The reason this method is so useful is because one only needs to calculate the Lagrangian displacements once, at redshift $z = 0$, and scale them by the appropriate derivatives of the growth factor. In fact, it is common practice in many N-Body simulations to

use 2LPT to generate the initial positions of the particles at a suitably high redshift, where the particle displacements more linear and 2LPT more accurate.

In L-PICOLA, Eq. 2.18 is solved as a whole (as opposed to evaluating Ψ_{res} individually) by discretising the operator 'T' using the Kick-Drift-Kick algorithm (Quinn et al., 1997), such that at each iteration the velocity and position of each particle is updated based on the gravitational potential Φ *and* the stored 2LPT displacements.

The equations for updating the particle positions and velocities can be modified to solve the COLA EOM in the following way

$$\boldsymbol{v}_{i+1/2} = \boldsymbol{v}_{i-1/2} - T[\boldsymbol{\Psi}_{res}]\Delta a_1, \tag{2.20}$$

$$\boldsymbol{r}_{i+1} = \boldsymbol{r}_i + \boldsymbol{v}_{i+1/2}\Delta a_2 + \Delta D_1 \boldsymbol{\Psi}_1 + \Delta D_2 \boldsymbol{\Psi}_2 \tag{2.21}$$

Here $\Delta D = D_{i+1} - D_i$ denotes the change in the first and second order growth factors over the timestep. The modified Kick-Drift-Kick equations are derived under the condition that, for Eq. 2.18 to be valid, the displacements felt by each particle must be computed in the 2LPT reference frame. In other words, the acceleration each particle feels due to the gravitational potential must be modified, and the 2LPT contribution to the acceleration removed. The new gravitational potential is then, by design, $T[\Psi_{res}]$. The exact procedure used to calculate $\nabla\Phi$ is not important and, as such, any code that updates the particle velocities and positions iteratively based on the gravitational potential, i.e., a Tree-PM code, can be modified in the above way to include the COLA mechanism.

An important point of note in enforcing the change of reference frame is that particle velocities at the beginning of the simulation, after creation of the 2LPT initial conditions but before iterating, must be identically 0. At this point the velocity a particle has is exactly equal to the velocity of the reference frame the particles are being moved to. However when the particles are output at the end of the simulation the required quantity is the particle velocities in *Eulerian* coordinates. This means that the initial particle velocities must be removed and stored at the beginning of the timestepping and then added back on at the end of the simulation.

When implementing the modified COLA timestepping, the time intervals, Δa_i, for each timestep do not get changed explicitly and as such can remain the same as those presented in Quinn et al. (1997). However, Tassev et al. (2013) present a second, COLA specific, formulation which gives faster convergence, hence allowing for recovery of the evolved dark matter field to greater accuracy in fewer time steps. In their method,

$$\begin{aligned}
\Delta a_1 &= \frac{H_0}{nLPT} \frac{a_{i+1/2}^{nLPT} - a_{i-1/2}^{nLPT}}{a_i^{nLPT-1}}, \\
\Delta a_2 &= \frac{H_0}{a_{i+1/2}^{nLPT}} \int_{a_i}^{a_{i+1}} \frac{a^{nLPT-3}}{H(a)} da.
\end{aligned} \tag{2.22}$$

where they find the best results using a value $nLPT = 2.5$. As the choice of Δa_i is somewhat arbitrary L-PICOLA retains both methods as options. This choice (and $nLPT$) should be treated formally as an extra degree of freedom in the code. In fact, it is shown later that the value of $nLPT$ used in the code can affect the final shape of the power spectrum recovered from COLA due to the way different growing modes are emphasised by different values. This is pointed out in Tassev et al. (2013) and means that for a given set of simulation parameters one would ideally experiment to find the type of timestepping that recovers the required clustering in the fewest timesteps possible. This is demonstrated further in Section 2.6.

For the timestepping method presented here and adopted in L-PICOLA, the only remaining piece of the puzzle is the calculation of $T[\mathbf{\Psi}_{res}] = -T[D_1]\mathbf{\Psi}_1 - T[D_2]\mathbf{\Psi}_2 - \nabla\Phi$. As the ZA and 2LPT displacements have been stored only a method of evaluating $\nabla\Phi$ is needed. In L-PICOLA this is done using the Particle-Mesh algorithm, though could be done using a method such as the Tree-PM algorithm. The evaluation of $T[D_1]$ and $T[D_2]$ can be performed numerically for a given cosmological model very easily, although a suitable approximation for D_2 must be adopted. For flat cosmologies one could use Eq. 2.9, however in L-PICOLA the expression of Matsubara (1995) is adopted, which is also accurate for non-flat cosmologies.

$$D_2(a) = -D_1^2(a)\left[\frac{\Omega_m}{4} - \frac{\Omega_\Lambda}{2} - \frac{1}{U_{3/2}}\left(1 - \frac{3U_{5/2}}{2U_{3/2}}\right)\right], \tag{2.23}$$

where

$$U_\alpha(a) = \int_0^1 \frac{da}{(\Omega_m a^{-1} + \Omega_\Lambda a^2 + \Omega_k)^\alpha} \equiv a^{2\alpha-1}E^{2\alpha}(a)\int_0^a \frac{da'}{a'^{2\alpha}E^{2\alpha}(a')}. \tag{2.24}$$

$E(a)$ is defined as the Hubble parameter with the Hubble constant scaled out, $H(a) = H_0E^2(a)$, such that, for example, $U_{3/2}$ contains factors of $E(a)$ raised to the power of 3.

In the code itself, several factors of H_0 and a are scaled out, as is the factor of $5\Omega_m/2$ in the linear growth factor. Ultimately, the cosmological quantities of interest are

$$a^3E(a)\frac{dD_1}{da} = \frac{1}{E(a)}\left(\left[\int_0^1 \frac{da'}{a'^3E^3(a')}\right]^{-1} - \left[\frac{3\Omega_{m,0}}{2a} + \Omega_{k,0}\right]D_1(a)\right), \tag{2.25}$$

$$a^3E(a)\frac{dD_2}{da} = \frac{a^2E(a)D_1^2(a)}{4a}\left[-3\Omega_m(a) + \frac{1}{U_{3/2}^2(a)}\left(2 + 4U_{3/2}(a) - 3[2 + \Omega_m(a) - 2\Omega_\Lambda(a)]U_{5/2}(a)\right)\right], \tag{2.26}$$

$$a^3E(a)\frac{d}{da}\left[a^3E(a)\frac{dD_1}{da}\right] = \frac{3}{2}\Omega_{m,0}aD_1(a), \tag{2.27}$$

$$a^3E(a)\frac{d}{da}\left[a^3E(a)\frac{dD_2}{da}\right] = \frac{3}{2}\Omega_{m,0}a(D_2(a) - D_1^2(a)). \tag{2.28}$$

2.2 A Lightcone-enabled Parallel Implementation of COLA (L-PICOLA)

As suggested by the name, L-PICOLA is a parallel implementation of the COLA method described in the previous section. It has been designed to be 'stand alone', in the sense that it can generate a dark matter realisation based solely on a small number of user defined parameters. This includes preparing the initial linear dark matter power spectrum, generating an initial particle distribution with k-space distribution that matches this power spectrum, and evolving the dark matter field over a series on user specified timesteps until some final redshift is reached. At any point in the simulation the particle position and velocities can be output, allowing the user to capture the dark matter field across a variety of epochs in a single simulation.

In order to make L-PICOLA as useful as possible several options have been implemented that modify how L-PICOLA is built at compile time. On top of allowing variations in output format and memory/speed balancing, the code can also be used to create (and then evolve) initial particle distributions containing primordial non-Gaussianity. Another significant improvement, and one that will be extremely important for future large scale structure surveys, is the option to generate a lightcone simulation, which contains variable clustering as a function of distance from the observer, as opposed to a snapshot simulation at one fixed redshift. Although lightcone simulations can be reconstructed from a series of snapshots (Fosalba et al., 2013; Merson et al., 2013), L-PICOLA can produce lightcone simulations 'on-the-fly' in a short enough time to be suitable for generating significant numbers of mock galaxy catalogues. These additions will be detailed and tested in later sections.

Fig. 2.1 shows a simple step-by-step overview of how L-PICOLA works. The different coloured boxes highlight areas where the structure of the code actually changes depending on how it is compiled. The blue box shows where the different types of non-Gaussianity can be included. The red boxes show where significant algorithmic changes occur in the code if lightcone simulations are requested.

2.3 Parallelisation

This section details the steps that have been taken to parallelise the COLA method. All parallelisation in the code uses the Message Passing Interface (MPI) library[1]. See Pacheco (1997) for a comprehensive guide to the usage and syntax of MPI. The following subsections detail the three main parallel algorithms in the code: parallel Cloud-in-Cell

[1]This software package can be downloaded at `http://www.open-mpi.org/`

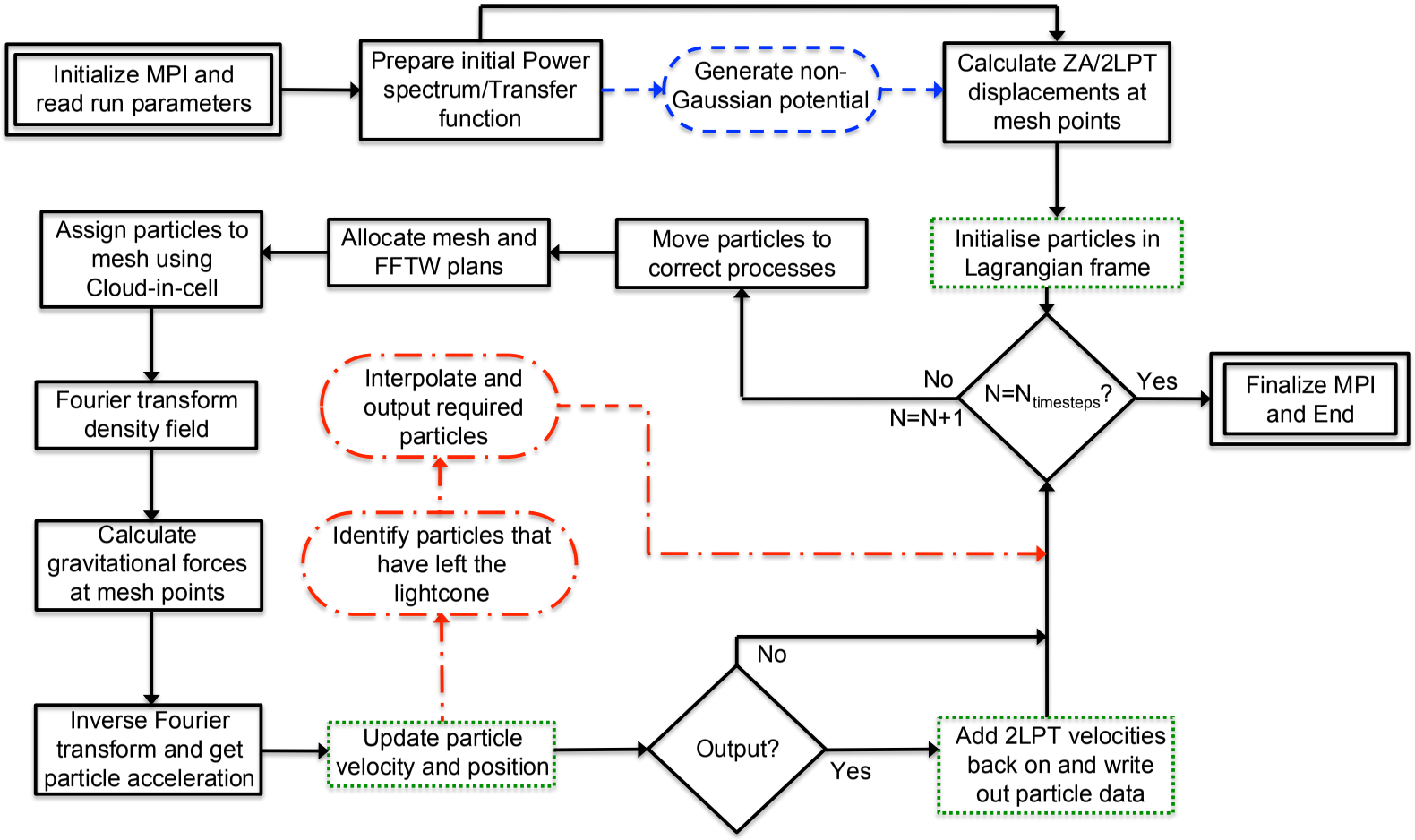


Figure 2.1: A flowchart detailing the steps L-PICOLA takes in generating a dark matter realisation from scratch. The green, dotted boxes indicate where the COLA algorithm is applied, differentiating L-PICOLA from a standard PM code. The blue, dashed box indicates where the inclusion of primordial non-Gaussianity changes the code structure. The red, dot-dash boxes highlight areas where the code differs depending on whether snapshot or lightcone simulations are being run.

interpolation, parallel FFT's and moving particles between processors.

2.3.1 Parallelisation Overview

Parallelisation of L-PICOLA has been performed with the goal that each processor can run a small section of the simulation whilst needing minimal knowledge of the state of the simulation as a whole. Both the mesh and particles have been split across processors in one direction. In this way each processor gets a planar portion of the mesh, and the particles associated with that portion. Extra care has been taken to balance the load on each processor as much as possible whilst adhering to the fact that each processor must have an integer number of mesh cells in the direction over which the full mesh has been split.

This process is enabled by use of the publicly available FFTW-MPI libraries, which also serve to perform the Fast Fourier Transforms when the mesh is split over different processors[2]. In a simulation utilising N_p processors and consisting of a cubic mesh of size N_m^3, each processor gets $(\lceil N_m/N_p \rceil)$ slices of the mesh where each slice consists of $N_m \times 2(N_m/2 + 1)$ cells. The extra $2N_m$ cells in each slice are required as buffer memory for the FFTW routines. Depending on the ratio of N_m to N_p this may give too many slices in total, so then the algorithm works backwards, removing slices until the total number of slices is equal to N_m.

The number of particles each processor has is related to the number of mesh cells on that processor. Each processor only requires knowledge of any particles that interact with its portion of the mesh. Hence, as the particles are originally spaced equally across the mesh cells, each processor initially holds N^3/N_m^2 particles multiplied by the number of slices it has.

2.3.2 Parallel Cloud-in-Cell

As each processor only contains particles which belong to the mesh cells it has, and the interpolation algorithm used to assign particles to the mesh 'rounds down', the density assignment step proceeds as per the standard Cloud-in-Cell interpolation method, except near the 'left-hand' edge of the processor. Here the density depends on particles on the preceding processor. Figure 2.2 shows a 2-D graphical representation of this problem.

In order to compensate for this an extra mesh slice is assigned to the 'right-hand' edge of each processor. This slice represents the leading slice on the neighbouring processor and by assigning the particles to these where appropriate and then transferring and adding the 'slices' to the appropriate processors, each portion of the mesh now contains an es-

[2]These are included in the FFTW package mentioned previously

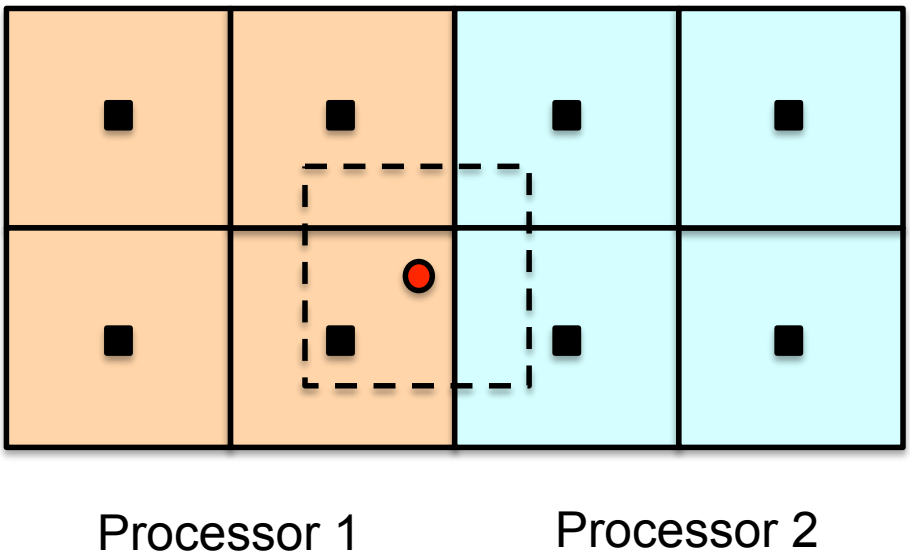


Figure 2.2: A visual representation of the 2-D Cloud-in-Cell algorithm. The particle (red point) is shared across the four nearest mesh cells with weight given by the percentage of the particle's 'cloud' (dashed line) that overlaps the mesh cell. In L-PICOLA these mesh cells may not be on the same processor as the particle. This is corrected by assigning extra slices of the mesh to the 'right-hand' edge of processor i which are then transferred and added to the left most slice on processor $i + 1$.

timate of the density which matches the estimate as if all the mesh were contained on a single processor.

It should also be noted that a reverse of this process must also be done after calculating the forces at each mesh point, as the displacement of a particle near the edge of a processor is reliant on the forces at the edge of the neighbouring processor.

2.3.3 Parallel FFT's

To take the Fourier transform of the mesh once it is split over many processors the parallel FFTW-MPI routines are used, which are available alongside the aforementioned FFTW libraries. This is intimately linked to the way in which the particles and mesh are actually split over processors and routines are provided in this distribution that enable this split to be performed in the first place.

The FFTW routines use a series of collective MPI communications to transpose the mesh and perform a multi-dimensional real-to-complex discrete Fourier transformation of the density, assigning the correct part of the transformed mesh to each processor. In terms of implementing this: First, the particles and mesh to be partitioned in a way that is compatible with the FFTW routines; Second, an FFTW 'plan' must be created for the arrays that are being transformed; and finally the Fourier transformation itself must be performed once the required quantity at each mesh point has been computed. The FFTW libraries perform all MPI communications and operations internally.

2.3.4 Moving Particles

One final modification to the Particle-Mesh algorithm is to compensate for the fact that, over the course of the simulation, particles may move outside the physical area contained on each processor. Their position may now correspond to a portion of the mesh that the processor in question does not have. As such, after each timestep the code checks to see which particles have moved outside the processor boundaries and then moves them to the correct processor. This is made particularly important as the COLA method converges in very few timesteps, meaning the particles can move large distances in the space of a single timestep.

In the case where there is a high particle density or small physical volume assigned to each processor, a single particle can jump across several neighbouring processors in a single timestep. So, when moving the particles, the code iterates over the maximum number of processors any single particle has jumped across. However, the number of particles that need to be moved is unknown *a priori* and so to be conservative and make sure that the buffer memory set aside for the transfer is not overloaded, not all the particles that are moving are transferred simultaneously (i.e. via a collective MPI-Alltoall command).

Rather, all the particles that have moved from processor N to N$\pm$1 are moved first then all the particles that have moved from processor N to N$\pm$2 are transferred. Although, in principle, this requires iterating over the particles on processor N multiple times, in practice the majority of cases have no particles moving to any processors beyond N$\pm$1 and so only one iteration is required.

As the simulation progresses the particles will not remain homogeneously spread over the processors, so additional buffer memory is assigned to each processor to hold any extra particles it acquires. This is utilised during the moving of the particles and all particles a processor receives are stored in this buffer. However, in order to make sure this buffer is not filled too quickly the fact that each processor is likely to lose some particles is also made use of. When a particle is identified as having left a particular processor the particle is moved into temporary memory and the gap is filled with a particle from the end of the processor's main particle memory. In this way all remaining particles are collected together before moving the new particles across, ensuring a contiguous, compact particle structure. This is shown in Figure 2.3.

2.4 Generating Initial Conditions

In order to allow L-PICOLA to run a simulation from scratch an initial conditions generator has been built into the code. This also means that the first and second order Lagrangian displacements for each particle can be stored as they are calculated rather than having to assume some initial positions for the particles and reconstruct their displacements. The initial conditions generation is based on the latest version of the parallelised 2LPTic code[3] (Scoccimarro, 1998; Scoccimarro et al., 2012), with some modifications to allow a more seamless combination of the two codes, especially in terms of parallelisation.

The initial conditions are generated using a linear matter power spectrum or transfer function for the simulation cosmology at redshift zero. This can either be read in, i.e., generated by an external program such as CAMB, or a transfer function can be generated using the model of Eisenstein & Hu (1998). The power spectrum may be also tilted and scaled by the values of n_s and σ_8 that are given to the code. Gaussian density fluctuations are drawn from the input power spectrum on a k-space grid using the same method as for generating a Gaussian realisation in Section 2.1.1. These are then scaled back to the redshift of the initial conditions via the linear growth factor. The initial, linear density perturbation are used as the basis for the first and second order LPT displacements.

For compatibility with L-PICOLA support for warm dark matter has been removed from the 2LPT initial conditions generator provided by Scoccimarro (1998), though this

[3]A parallelised version of the code including primordial non-Gaussianity can be found at `http://cosmo.nyu.edu/roman/2LPT/`.

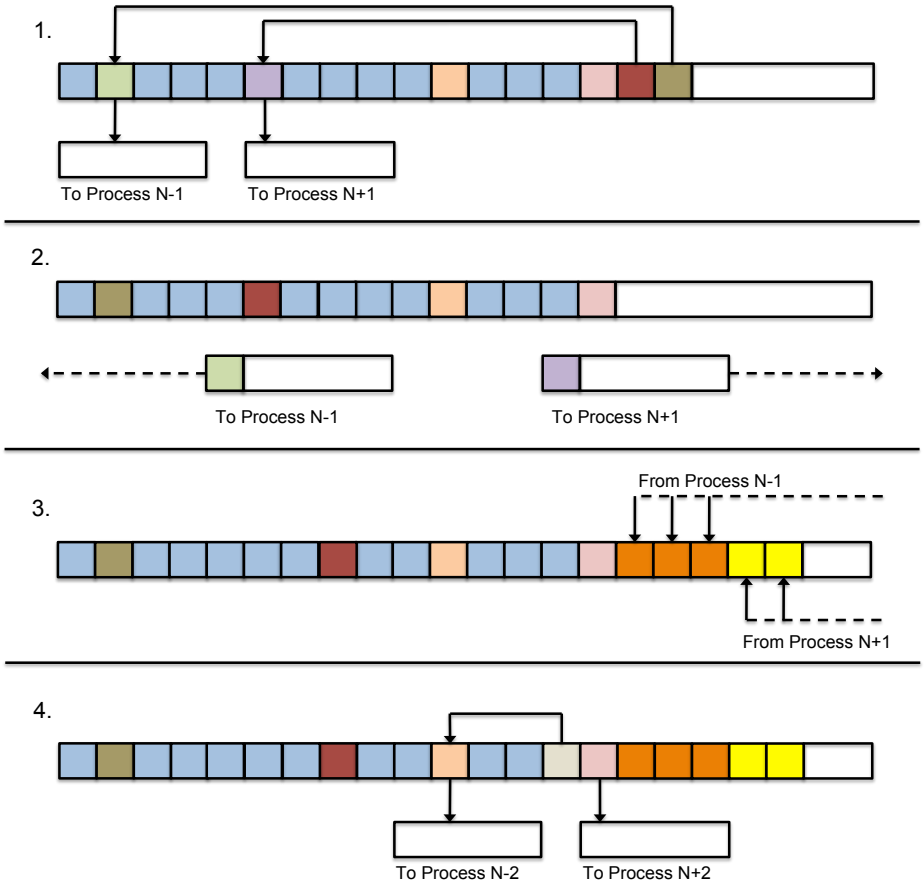


Figure 2.3: A four stage 'memory schematic' of how L-PICOLA moves particles between processors in between timesteps, conserving as much memory as possible. First those particles which need moving to the neighbouring processors are identified and moved to a temporary buffer. Then particles are moved from the end of the particle structure to overwrite the particles that the processor no longer needs to keep. Finally a send and receive operation is performed, sending the particles in the buffer to the neighbouring processors and receiving particles from those processors into the end of the particle structure. This algorithm is repeated up to the maximum number of processors a particle has moved across.

improvement could easily be added in the future. The particles are initially placed uniformly, in a grid pattern throughout the simulation volume, so rather than creating the particles at this stage, memory is conserved by only generating the 2LPT displacements at these points and creating the particles themselves just before timestepping begins.

Because of the addition of the initial conditions generator, L-PICOLA can be used very effectively to create the initial conditions for other N-Body simulations, as well as evolving the dark matter field itself. In fact in a single run both the initial conditions and the evolved field can be output at any number of redshifts between the redshift of the initial conditions and the final redshift, which allows easy comparison between L-PICOLA and other N-Body codes.

A final point is that because the 2LPT section is based on the latest version of the 2LPTic code, L-PICOLA is able to generate, and then evolve, initial conditions with local, equilateral, orthogonal or generic primordial non-Gaussianity. Local, equilateral and orthogonal non-Gaussianity can be added simply by specifying the appropriate option before compilation and providing a value for f_{NL}. Primordial non-Gaussianity for any generic bispectrum configuration can be generated using a user-defined input kernel, following the formalism in Appendix A of Scoccimarro et al. (2012), which defines a general operator for the non-Gaussian potential based on the input power spectrum.

2.5 Simulating Lightcones

Another feature contained within L-PICOLA, which will be very useful for future large scale structure surveys, is the ability to generate lightcone simulations in a single run, as opposed to running a large number of snapshots and piecing them together afterwards.

Snapshot simulations, generated at some effective redshift, have been widely used in the past to calculate the covariance matrix and perform systematic tests on data (e.g., Manera et al. 2013, 2015). However, as future surveys begin to cover larger and larger cosmological volumes with high completeness across all redshift ranges it is no longer good enough to produce a suite of simulations at one redshift. Lightcone simulations mimic the observed clustering as a function of redshift and so introduce a redshift dependence into the covariance matrix. On top of this, once a full redshift range has been simulated one can apply identical selection functions, such as redshift cuts, to the mock galaxy catalogues and the data. In the case where measurements at multiple effective redshifts are made with a single sample, e.g., the DES and Euclid surveys (The Dark Energy Survey Collaboration, 2005; Laureijs et al., 2011), fewer simulations may be needed in total, especially if multiple runs would be required to produce the snapshots at multiple redshifts.

Figure 2.4 demonstrates the effect of simulating a lightcone using the power spec-

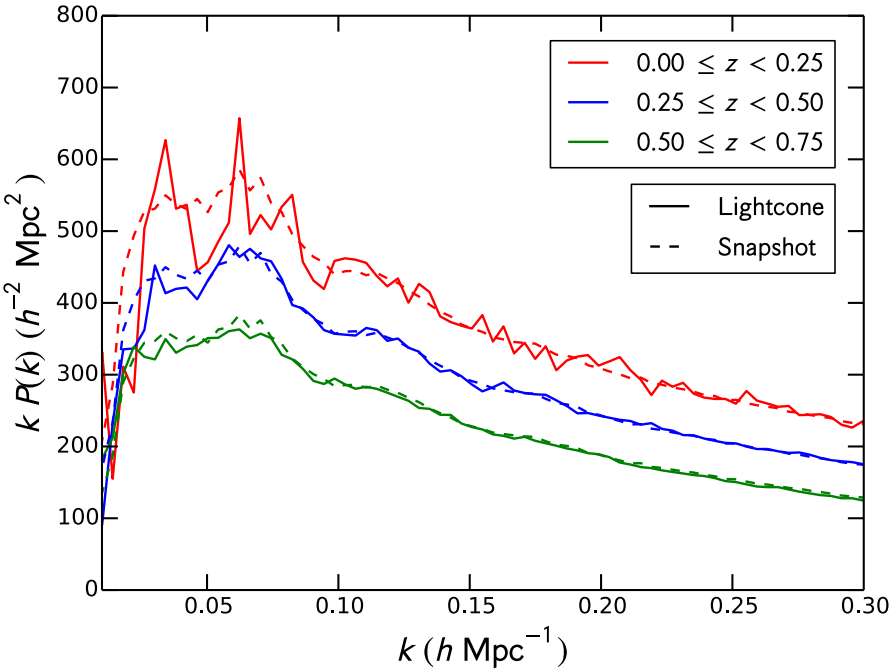


Figure 2.4: The power spectra, measured using the estimator of Feldman et al. (1994), of different redshifts slices within the same L-PICOLA lightcone simulation (solid). This is compared to a snapshot simulation at the effective redshift of the slice. As expected, the clustering is much stronger (and the power spectrum amplitude much higher) at lower redshifts and there is good agreement between the snapshot and lightcone power spectra.

trum. These were made by simulating a $(2\,h^{-1}\,\mathrm{Gpc})^3$ box with 512^3 particles. Placing the observer at (0,0,0) and using a flat, $\Omega_{m,0} = 0.25$ cosmology (all other parameters match the fiducial cosmology used within this chapter), allows a simulation of this size to simulate an eighth of the full-sky out to a maximum redshift of 0.75. The power spectrum is then calculated using the method of Feldman et al. (1994) for three redshift slices between 0.0, 0.25, 0.5 and 0.75, using a random, unclustered catalogue to capture the window function.

As expected, there is significant evolution of the clustering as a function of redshift that would not be captured in a single snapshot simulation. The overall clustering amplitude increases for lower redshifts with additional non-linear evolution on small scales at later times. To further compare the clustering of this lightcone simulation with the expected clustering, this is overlayed with the power spectra from a snapshot simulation at the effective redshift of each lightcone slice.

There is good agreement on all scales between the snapshot and lightcone power spectra for each of the redshift slices. The finite volume of the lightcone region causes noise on the largest scales, especially for the lowest volume slice, however the redshift-

dependent amplitude is captured very well within a single lightcone simulation.

The following subsections will provide a detailed description of how lightcone simulations are produced in L-PICOLA, test the accuracy of the implementation and also looking at how the simulation volume can be 'replicated' to fill the full lightcone during run-time.

2.5.1 Building Lightcone Simulations

Simulating the past lightcone requires knowledge of the properties of each particle in the simulation at the moment when it leaves a lightcone shrinking towards the observer. As has been done in several studies (Fosalba et al., 2013; Merson et al., 2013), these particle properties can be interpolated using a set of snapshot simulations. However, this requires significant post-processing and more storage space than generating a lightcone simulation at run-time. As such, in order to provide a useful tool for future cosmology surveys, this feature has been implemented into L-PICOLA.

The procedure works as follows: the user specifies an initial redshift, at which to begin the simulation, and an origin, the point at which the observer sits. Each of the output redshifts is then used to set up the timesteps used in the simulation, with the first output denoting the point at which to start the lightcone and the final output corresponding to the final redshift of the simulation. Any additional redshifts in between these two can be used to set up variable timestep sizes. If one imagines the lightcone as shrinking towards the origin as the simulation progresses, then for every timestep between these two redshifts only those particles that have left the lightcone are output. This is shown pictorally in Figure 2.5.

Mathematically, it is simple to identify whether the particle should be output between timesteps i and $i+1$ by looking for particles which satisfy both

$$R_{p,i} \leq R_{L,i} \tag{2.29}$$

and

$$R_{p,i+1} > R_{L,i+1}, \tag{2.30}$$

where $R_{p,i}$ is the comoving distance between the particle and the lightcone origin at scale factor $a_i = 1/(1+z_i)$ and $R_{L,i}$ is the comoving radius of the lightcone at this time.

Ideally a given particle should be output at the exact moment it satisfies the equation,

$$R_p(a_L) = R_L(a_L). \tag{2.31}$$

From the COLA method

$$R_p^2(a_L) = |\boldsymbol{r}_i - \boldsymbol{r}_0 + \boldsymbol{v}_{i+1/2}\Delta a_2 + \Delta D_1 \boldsymbol{\Psi}_1 + \Delta D_2 \boldsymbol{\Psi}_2|^2 \tag{2.32}$$

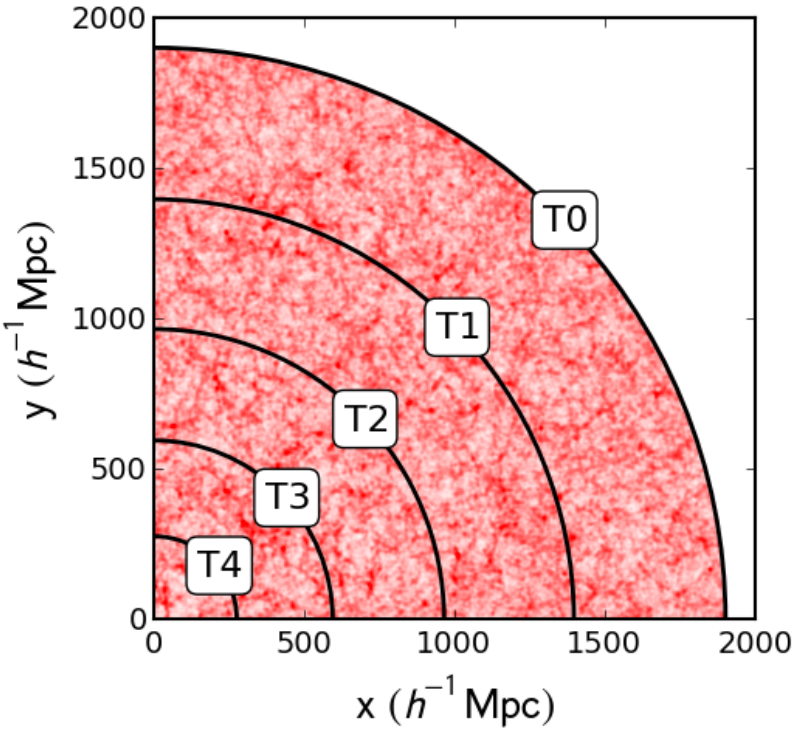


Figure 2.5: A $50\,h^{-1}\,\mathrm{Mpc}$ slice of a L-PICOLA dark matter field simulated on the past lightcone with an observer situated at the origin. As the lightcone shrinks with each timestep (the lightcone radius is denoted by the black lines) only the particles that have left the lightcone that timestep are output, with their position interpolated to the exact point at which they left. This means that the particles shown in the diagram were output in stages with the particles between lines T0 and T1 output first. Between output stages the particles evolve as normal, resulting in clustering that is dependent on the distance from the observer.

where $\boldsymbol{r}_0$ is the position of the lightcone origin and the 'Δ' terms are dependent on the value of a_L. The comoving lightcone radius at a_L is simply the comoving distance

$$R_L(a_L) = c\int_{a_L}^{1} \frac{da}{a^2 H(a)}. \tag{2.33}$$

Equating these allows a_L to be solved for. Once this is known it is simple to calculate the properties of each particle that needs outputting, as identified by Eqs. 2.29 and 2.30. However, this equation cannot be solved analytically and so requires the code to numerically solve it for each individual particle that it wishes to output. This would be prohibitively time-consuming. Instead an approximate solution is calculated by linearly interpolating both the lightcone radius and the particle position between the times a_i and a_{i+1}. Substituting the linear interpolation into Eq. 2.31 and rearranging

$$a_L \approx a_i + \frac{(a_{i+1} - a_i)(R_{L,i} - R_{p,i})}{(R_{p,i+1} - R_{p,i}) - (R_{L,i+1} - R_{L,i})}. \tag{2.34}$$

This is trivial to calculate as $R_{p,i}$ and $R_{p,i+1}$ already need to be known in order to update the particle during timestepping anyway, and $R_{L,i}$ and $R_{L,i+1}$ are needed to identify which particles have left the lightcone in the first place. In fact the whole procedure can be performed with minimal extra runtime, by simply modifying the 'Drift' part of the code. The only extra computations are to check the particle's new position against the lightcone and interpolate if necessary. Once the exact time the particle left the lightcone is known, the particle's position can be updated to the position it had when it left the lightcone using Eq. 2.21 before it is subsequently output.

In L-PICOLA lightcone simulations the velocity is not interpolated. The velocity at time $a_{i+1/2}$ is used instead. This choice was made as it mimics the inherent assumption of the Kick-Drift-Kick method, that the velocity is constant between a_i and a_{i+1}. To properly interpolate the velocity in the same way as the particle position would require the code to evaluate the velocity for each particle at times a_i and a_{i+1} which in turn would require the particle density to be measured at half timestep intervals. One could also imagine assuming that the non-linear velocity is constant and interpolating the ZA and 2LPT velocities (which must be added back on before outputting to move back to the correct reference frame). However, as shown below it was found that the assumption of constant velocity between a_i and a_{i+1} is a reasonable one.

To test the numerical interpolation against the analytic expectations, and provide a graphical representation of the particle positions and velocities output during lightcone simulations, particles output during the final timestep of a lightcone simulation are compared to the same particles output from snapshot simulations evaluated at the beginning and end of that timestep (the corresponding redshifts are $z = 0.0$ and $z = 0.09375$ in this case). The particles are matched based on a unique identification number which is

assigned when the particles are created and as such is consistent between the three simulations.

For both the particle positions and velocities the quantity of interest is the difference between the lightcone properties and the properties of the $z = 0.0$ snapshot, normalised by the same difference between the $z = 0.09375$ and $z = 0.0$ snapshots. This is plotted in Figure 2.6 as a function of the distance from the observer (also normalised using the comoving distance to $z = 0.09375$). If the particle positions were interpolated after runtime using the two snapshots one would expect the particles to lie exactly on the diagonal in Figure 2.6. In L-PICOLA the particle postions interpolated *during* the simulation also lie close to the diagonal, which validates the accuracy of the numerical interpolation. The small scatter in both of these plots is due to floating-point errors and the normalisation in the particle positions. Particles that do not move much between the two snapshots will have a normalisation close to zero, which in turn makes the choice of plotting statistic non-optimal. The particle velocities show no trend as a function of distance to the observer or when they were output. In this case the velocities in each direction are all situated close to the mid point between the two simulations . This validates the Kick-Drift-Kick assumption, that the velocities evolve approximately linearly between two timesteps, such that the velocity at time $a_{i+1/2}$ is half way between that at time a_i and a_{i+1}, although there is some scatter and offset due to the true non-linear nature of the velocity.

Interpolation Accuracy

On top of comparing the numerical interpolation during runtime to the analytic interpolation between two snapshots, the validity of the assumption that linear interpolation can be used between two timesteps is tested. As mentioned previously, the exact time the particle leaves the lightcone is given by numerically solving Eq. 2.31, but solving this for each particle is extremely time consuming and a solution is computed using linear interpolation instead. To test this, the exact solution is computed for a subset of the particles in the $L = 2\,h^{-1}\,\mathrm{Gpc}, N = 512^3$ simulation and compared to the approximate solution, $a_{L,interp}$. This is shown in Figure 2.7. As shown in this figure, linear interpolation slightly overestimates the value of a_L, with a common trend across all timesteps, however this effect is less than 0.5% across all times for this simulation. The two solutions agree almost perfectly close to the timestep boundaries, denoted by the dashed vertical lines. This is because the particle positions and lightcone radius are known exactly at these points. Further away from the timestep boundaries, inaccuracies are introduced as the assumption that the particle position and lightcone radius are linear functions of the scale factor is less accurate.

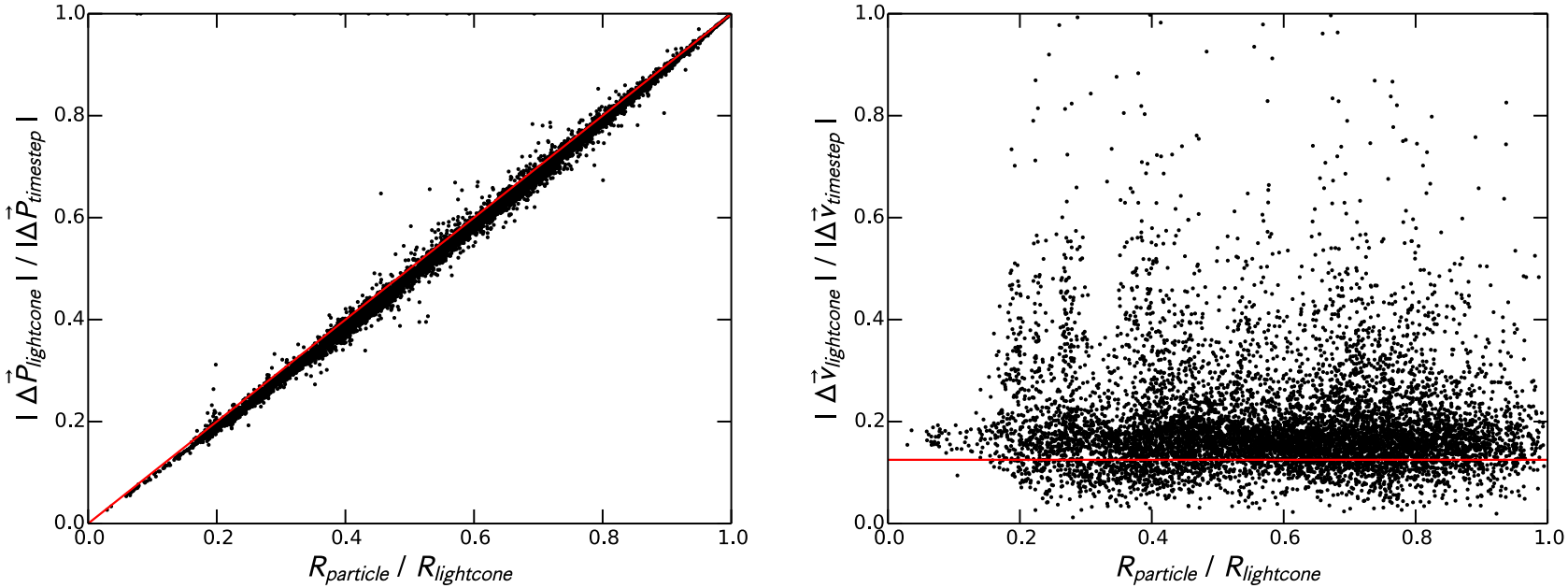


Figure 2.6: The difference between the lightcone and snapshot positions (left) and velocities (right) of particles output between $z = 0.0$ and $z = 0.09375$ as a function of the distance to the observer, which is equivalent to the output time. The exact quantity plotted is the magnitude of the difference vector between the lightcone and $z = 0.0$ snapshot statistics, normalised by the difference between the $z = 0.0$ and $z = 0.09375$ snapshots. The solid red line shows the expected trend based on the fact that particles are output at the exact time they exit the lightcone, but the velocity is not interpolated. For the latter the expectation is, for each direction, $v_{a+1/2} \approx (v_{a+1} + v_a)/2$.

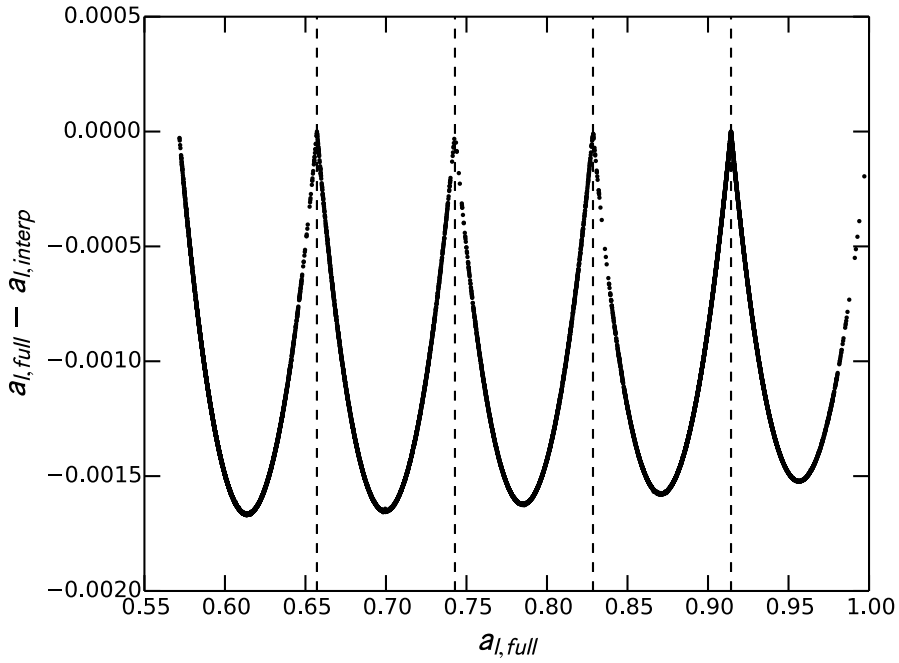


Figure 2.7: A plot showing the accuracy of using linear interpolation to get the time a particle leaves the lightcone. For a subset of particles, the difference between the full numerical solution of a_L and the value recovered using Eq. 2.34 is plotted as a function of the true scale factor. The dashed lines show the scale factor at which the code evaluates the timesteps of the simulations, and hence knows the exact positions of the particles and the lightcone radius.

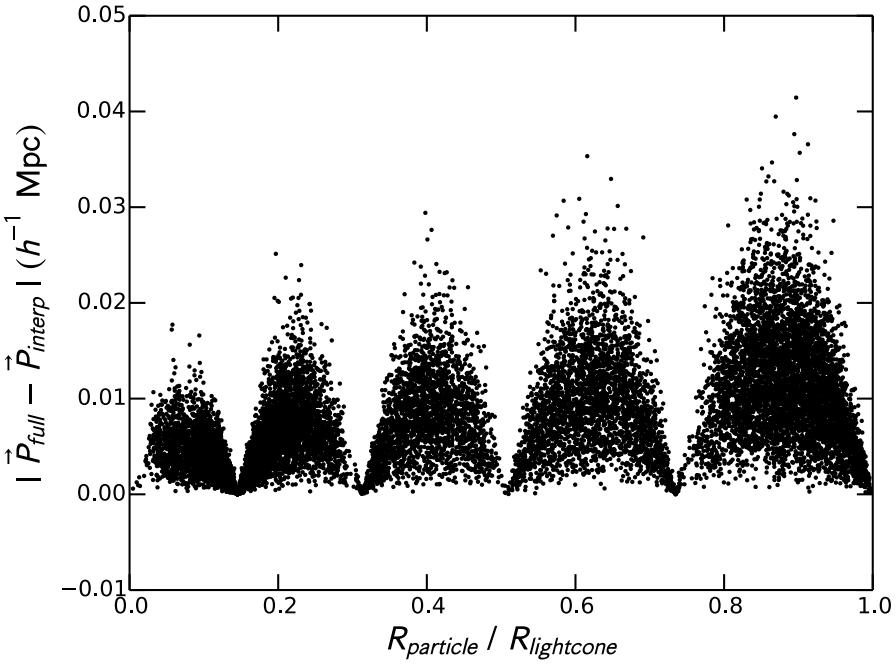


Figure 2.8: A plot of the difference between the positions of a subset of particles when using the full numerical solution of a_L and those recovered using Eq. 2.34, as a function of the distance from the observer, normalised by the maximum lightcone radius of the simulation. The plotted quantity is the magnitude of the difference vector between the two methods. There is good agreement, with a maximum difference of $\sim 50h^{-1}\,\mathrm{kpc}$ across all scales.

The reliability of the linear interpolation can be further quantified by looking at the positions of the particles output in both these simulations. This is shown in Figure 2.8, which plots the difference in particle position (the magnitude of the difference vector) as a function of the distance between the particle and the observer, normalised to the maximum lightcone radius for the simulation. The linear interpolation is indeed very accurate, and even at large radii, where the comoving distance between timesteps is largest, the particle positions are equivalent to within $0.05\,h^{-1}\,\mathrm{Mpc}$. This is well below the mesh scale of this simulation, and is subdominant compared to the errors caused by the finite mesh size and the large timesteps.

2.5.2 Replicates

The lightcone implementation within L-PICOLA can also account for the fact that lightcones built from snapshot simulations often replicate the simulation output to reach the desired redshift. L-PICOLA has the ability to replicate the box as many times as required in each direction during runtime. This is done by simply modifying the position of each

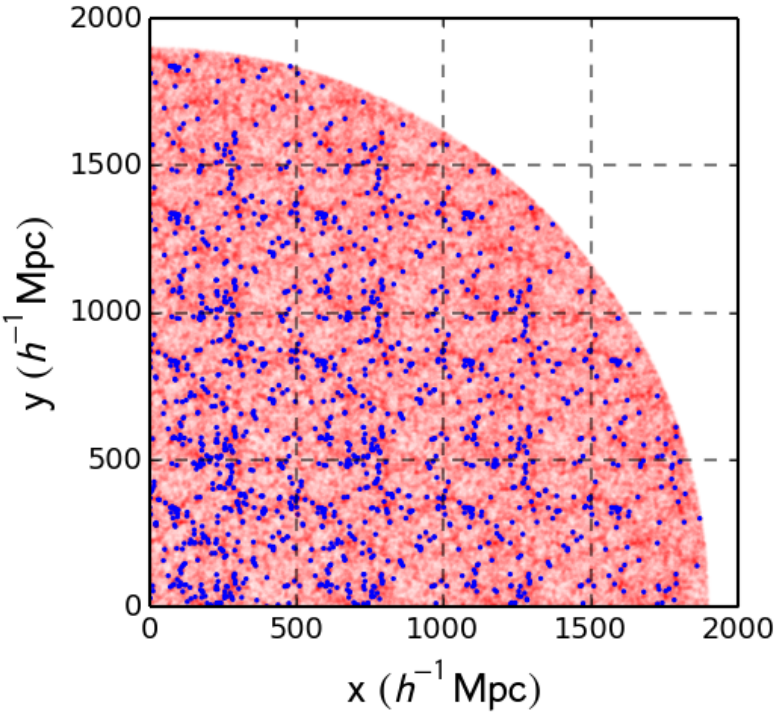


Figure 2.9: An L-PICOLA lightcone simulation showing obvious replicates. In this rather extreme case a similar simulation to that shown in Figure 2.5 is run, but using 64 times less particles in a volume 64 times smaller. The box is then replicated 64 times at runtime as shown by the dashed lines (only 1 replicate is shown in the z direction). To aid visualisation the halos recovered from this simulation using a Friends-of-Friends algorithm are also plotted in blue.

particle as if it were in a simulation box centred at some other location. In this way the user can build up large cosmological volumes whilst still retaining a reasonable mass and force resolution. However, it is important to note that this can have undesired effects on the power spectrum and covariance matrix calculated from the *full* replicated simulation volume, which will be detailed subsequently. Figure 2.9 highlights the replication process. This is done by running a similar lightcone to that used in Figure 2.5, however the actual simulation contains 64 times fewer particles in a volume 64 times smaller and is replicated 64 times. The simulation contains the full volume and mass resolution required but the CPU and memory requirements are much smaller. To help identify the replication the Friends-of-Friends algorithm (Davis et al., 1985) has been used to group the particles into halos and the centre-of-mass position of each halo has been plotted. This results in obvious points where the same halo is reproduced after more particles have accreted onto that halo, and the halo has evolved in time.

Effects of Replication on the Power Spectrum

The downside of the replication procedure is that in repeating the same structures there are fewer independent modes to sample compared to what would be expected from an unreplicated simulation of the same volume. Rather the same modes are being sampled multiple times. This affects both the power spectrum and the covariance matrix. The effects of replication are tested using a set of 500 lightcone simulations, containing 512^3 particles in a box of edge length $1024\,h^{-1}\,\mathrm{Mpc}$. This is compared to another set of 500 simulations with 256^3 particle in a $(512\,h^{-1}\,\mathrm{Mpc})^3$ box, which is then replicated 8 times. The power spectra for both sets are calculated using the method of Feldman et al. (1994), in bins of $\Delta k = 0.008h\,\mathrm{Mpc}^{-1}$, estimating the expected overdensity from the total number of simulation particles and the box volume. This works for the lightcone simulations as the maximum lightcone radius is larger than the diagonal length of the cubic box, such that the simulation still fills the volume.

The average power spectra are shown in Figure 2.10, where the errors come from the diagonal elements of the covariance matrix and are those for a single realisation. As the simulations are periodic in nature one would expect the power spectra for the two box sizes to be almost identical except for the fact that the larger simulation volume has a greater effective redshift and hence a power spectrum with lower amplitude and less non-linear evolution. We see that this holds true for the lightcone simulations used here, and that the difference in the replicated and unreplicated 512^3 simulations is, at least on linear scales, equal to the difference in the linear growth factor between the effective redshifts of the two sets of simulations (the effective redshifts and normalised linear growth factors are $z_{eff} = 0.17, D_1(z_{eff}) = 0.91$ and $z_{eff} = 0.36, D_1(z_{eff}) = 0.82$ for the small and large boxes respectively).

However, in order to produce the replicated power spectrum, it is necessary to correct for the replication procedure. When a simulation is replicated, the fundamental mode of the simulation is changed but without adding any additional information, either in the number of independent modes sampled, or on scales beyond the box size of the unreplicated simulation. This in turn creates ringing on the order of the unreplicated box size. This can demonstrated using a simple toy model.

Figure 2.11 shows a small 2×2 overdensity field before and after taking the discrete Fourier transform. Then, if the 2×2 overdensity field is replicated 4 times and a discrete Fourier transform performed, the resultant Fourier components are assigned to a grid 4 times larger than for the unreplicated field as the fundamental mode of the simulation should be twice as small. However replication has not added any information beyond that contained in the original 2×2 grid and as such every other component of the Fourier transformed replicated field is zero, creating ringing within the power spectrum.

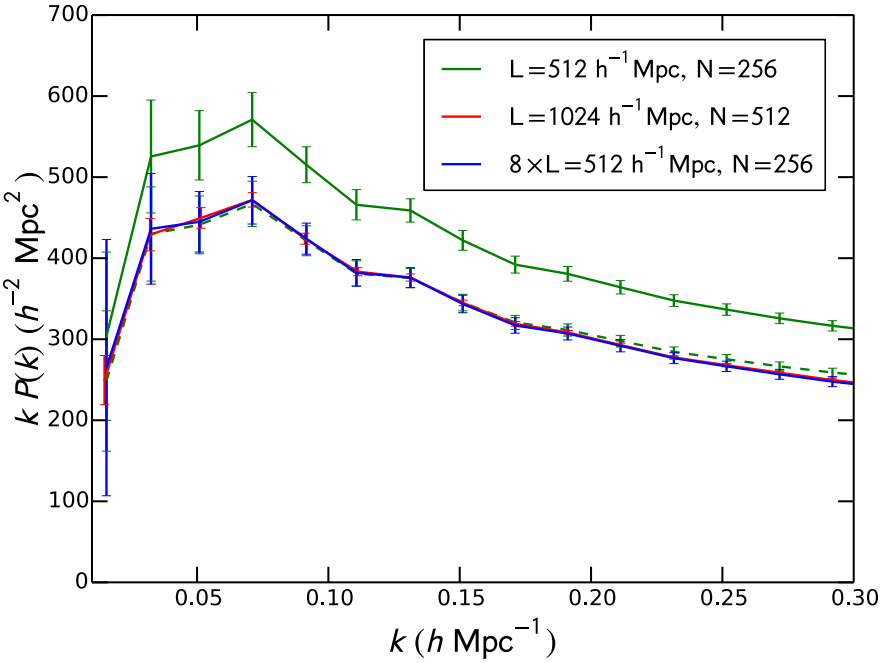


Figure 2.10: Plots showing the effect of replication on the estimated power spectrum using sets of 512^3 particle lightcone simulations in a $1024\,h^{-1}\,\mathrm{Mpc}$ box and 256^3 particle lightcone simulations in a $512\,h^{-1}\,\mathrm{Mpc}$ box. The lines correspond to the average power spectra from 500 independent realisations and the errors are those on a single realisation calculated from the diagonal of the covariance matrix constructed from the 500 realisations. The blue line represents the average power spectrum when the 256^3 particle simulation is replicated 8 times so that it has the same volume and number of particles as the larger simulation, and as expected is virtually indistinguishable from the large, unreplicated simulation. The larger amplitude of the green line is due to the lower effective redshift of the smaller box and this amplitude difference can be scaled out at linear scales using the linear growth factor as shown by the dashed green line.

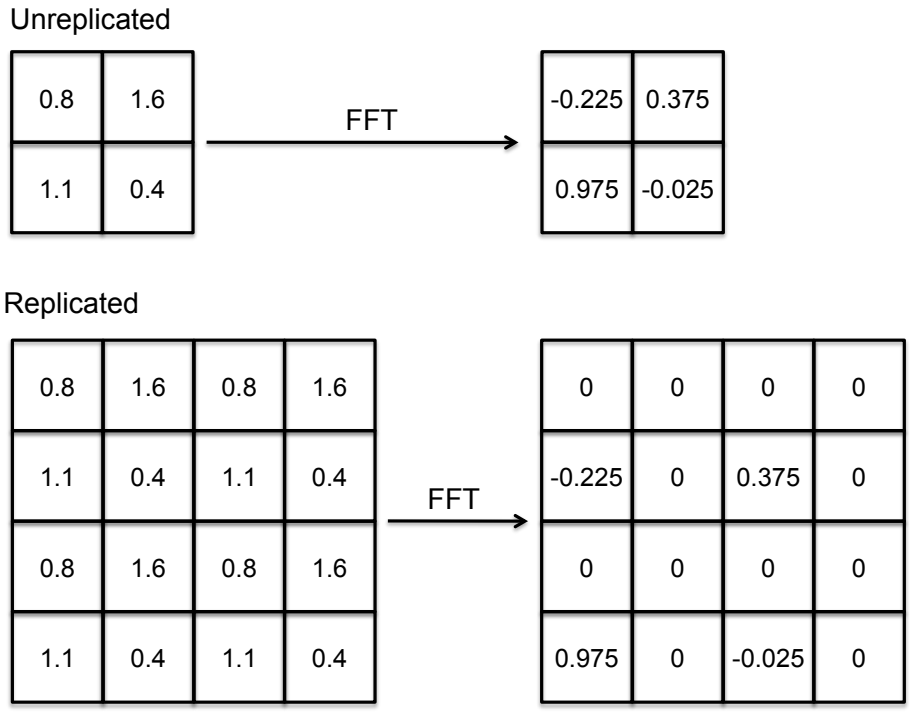


Figure 2.11: A toy model demonstrating how replication of the simulation volume can create ringing in the power spectrum. Replicating the simulation does not add any information below the fundamental mode of the unreplicated simulation. The lack of additional information then creates '0' elements in the Fourier transformed overdensity grid, which in turn creates ringing in the power spectrum.

This also highlights the correction that can be performed to remove this affect. After Fourier transforming the replicated overdensity field one need simply remove the zero components and place the remaining non-zero components in the same size grid as that used for the unreplicated box, correcting for the differences in normalisation between the two fields. The power spectrum is then computed using this smaller grid. This removes the ringing on the order of the box size and returns the power spectrum as seen in Figure 2.10. It is important to note that this procedure still lacks the k-space resolution one would naively expect from a simulation box that is larger. Neither the replication method in L-PICOLA nor the correction for ringing adds in modes larger than the unreplicated box size (there are, however, methods that do do this, see e.g., Tormen & Bertschinger (1996); Cole (1997)

This is an important correction and one that should be used whenever a simulation is replicated. It is important to note however that it is believed that such a correction will only be necessary when looking at a portion of a replicated simulation with volume equal to or greater than the unreplicated simulation. For most practical applications, the unreplicated simulation would be much larger than that used here, and the lightcone simulations themselves would undergo significant post-processing, such as the application of a survey window function and cutting into redshift slices. In this case the volume of each redshift slice will most likely be less than the original unreplicated simulation volume and so no correction will be necessary.

Effects of Replication on the Covariance

Utilising the 500 realisations for both sets of simulations also allows for investigation into the effect of replication on the covariance matrix. This is shown in Figure 2.12. Assuming Gaussian covariance, i.e., Tegmark (1997) (see also Chapter 5) one would expect the covariance to scale as the inverse of the simulation volume. The two sets of unreplicated simulations show this behaviour, with the larger volume simulation having a covariance 8 times smaller than the smaller simulation, at least on linear scales. But, as with the power spectrum, artificially increasing the simulation volume by replication does not add in any extra unique modes and so does not increase the variance. This in turn means that the covariance matrix of the replicated simulation does not display the expected volume dependence.

Knowing the expected volume dependence, however, allows one to correct for this effect. This correction is shown in 2.12 as the dashed blue line. The corrected, replicated covariance agrees very well with the unreplicated covariance, however there are some residual differences on small scales. It is hypothesised that this arises due to the absence of modes larger than the unreplicated box size, which would otherwise couple with the

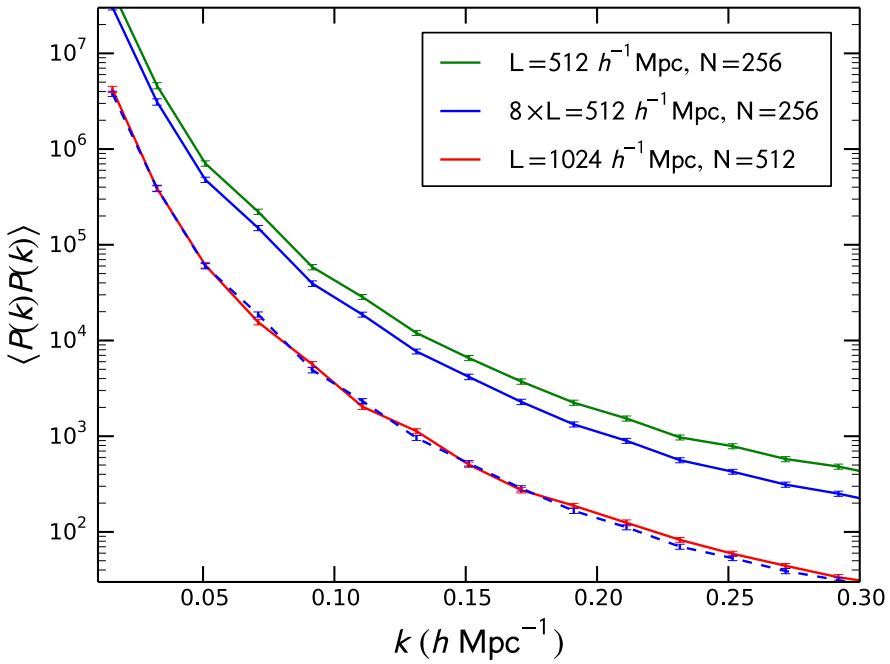


Figure 2.12: Plots showing the effect of replication on the diagonal elements of the power spectrum covariance matrix using sets of 512^3 particle lightcone simulations in a $1024\,h^{-1}\,\mathrm{Mpc}$ box and 256^3 particle lightcone simulations in a $512\,h^{-1}\,\mathrm{Mpc}$ box. The errors are derived from bootstrap resampling with replacement over the 500 realisations. The dashed line shows the covariance of the replicated simulations after dividing by the difference in volume between the two sets of unreplicated simulations.

small scale modes within the simulation and increase the small scale covariance. This coupling is referred to as the Super-Sample covariance by Takada & Hu (2013) and Li et al. (2014a), who also explore corrections for this effect that could be applied to replicated simulations. This phenomenon will be revisited in Chapter 5.

Like the power spectrum, most applications of L-PICOLA will involve some manipulation of the final simulation output, so it is not expected that this incorrect volume dependence will be present unless the comoving volume of the region being analysed is close to the unreplicated simulation size.

On the other hand, with this in mind, it is still recommended that for any usage of L-PICOLA involving replication of the simulation region, the effects on the power spectrum and covariance matrix are throughly tested. This could be done using a procedure similar to that shown here, comparing replicated and unreplicated simulations after applying any survey geometry and data analysis effects. Obviously replication will only be necessary if maintaining both the full volume and number density is unfeasible. As these effects arise due to the simulation volume rather than the particle number density one would be able to test this without simulating the full number of particles in the unreplicated volume.

Speeding Up Replication

In L-PICOLA, lightcone simulations are performed in such a way as to add no additional memory requirements to the run. However, the amount of time to drift the particles will increase in proportion to the number of replicates. In order to speed this up the code identifies, each timestep, which replicates are necessary to loop over. Any replicates that have all 8 vertices inside the lightcone at the end of the timestep will not have particles leaving the lightcone and so can be ignored for the current iteration. Furthermore, for replicates not fully inside the lightcone, the code calculates the shortest distance between the replicate and the origin by first calculating the distance to each face of the replicate then the shortest distance to each line segment on that face. If the shortest distance to the origin is larger than the lightcone radius then the replicate has completely exited the lightcone and will no longer be required for the duration of the simulation. Overall, this means that even if the simulation box is replicated N times in each direction the code will only need to look at a small fraction of the replicates ($\sim 1-2$ in each direction unless the simulation box is so small that the lightcone radius changes by more than the boxsize in a single timestep).

2.6 L-PICOLA Accuracy

This section compares the accuracy of L-PICOLA to a full N-Body simulation using the Tree-PM code GADGET-2 (Springel, 2005) and to the results returned using only 2LPT, which has been used to generate mock catalogues for the BOSS survey (Manera et al., 2013, 2015). In all cases simulations containing 1024^3 particles in a box of edge length $768\,h^{-1}\,\mathrm{Mpc}$ are compared, with the L-PICOLA simulations starting at an initial redshift of 9.0 and evolving to a final redshift of 0.0. The fiducial cosmology of this chapter is used along with a linear power spectrum calculated at redshift 0.0 from CAMB (Lewis et al., 2000; Howlett et al., 2012). In all cases, unless this choice itself is being tested, $N_{mesh} = N_{particles}$.

2.6.1 Two-point Clustering

The first test is how well L-PICOLA recovers the two-point clustering of the dark matter field compared to the N-Body simulation, which will be treated as the 'correct' solution. In all cases the power spectrum within the cubic simulations is calculated using the method of Feldman et al. (1994). Figure 2.13 shows the ratio of the power spectra recovered from the approximate simulations and from the GADGET-2 run. The plotted results are those recovered using 2LPT and L-PICOLA runs with 10 timesteps and 50 timesteps, and for a set of runs with the COLA modification turned off and the same numbers of timesteps. The act of turning the COLA method off reduces L-PICOLA to a standard Particle-Mesh code. Also plotted is the cross correlation, ρ, between the approximate dark matter field, δ, and the full non-linear field from the GADGET-2 run, δ_{NL}, defined as

$$\rho(k) = \frac{\langle \delta(\boldsymbol{k}) \delta^*_{NL}(\boldsymbol{k}) \rangle}{\langle |\delta(\boldsymbol{k})|^2 \rangle \langle |\delta_{NL}(\boldsymbol{k})|^2 \rangle} \tag{2.35}$$

The contribution from shot-noise is neglected here as for dark matter particles and the simulation specifications this will be very sub-dominant on all scales considered.

The COLA method creates a much better approximation of the full non-linear dark matter field than 2LPT and the Particle-Mesh algorithms alone for a small number of timesteps. The agreement between the COLA and N-Body fields is remarkable, with the power spectra agreeing to within 2% up to $k = 0.3\,h\,\mathrm{Mpc}^{-1}$, which covers all the scales currently used for BAO and RSD measurements. An 80% agreement remains even up to scales of $k = 1.0\,h\,\mathrm{Mpc}^{-1}$. This level of conformity is mirrored in the cross correlation, which for the COLA run remains above 98% for all scales plotted.

Further to this, where the cross-correlation is 1, this would not be expected to deviate between realisations. It is non-stochastic. As such where the cross-correlation is 1, the covariance of the L-PICOLA and GADGET-2 simulations is identical (at the level of noise

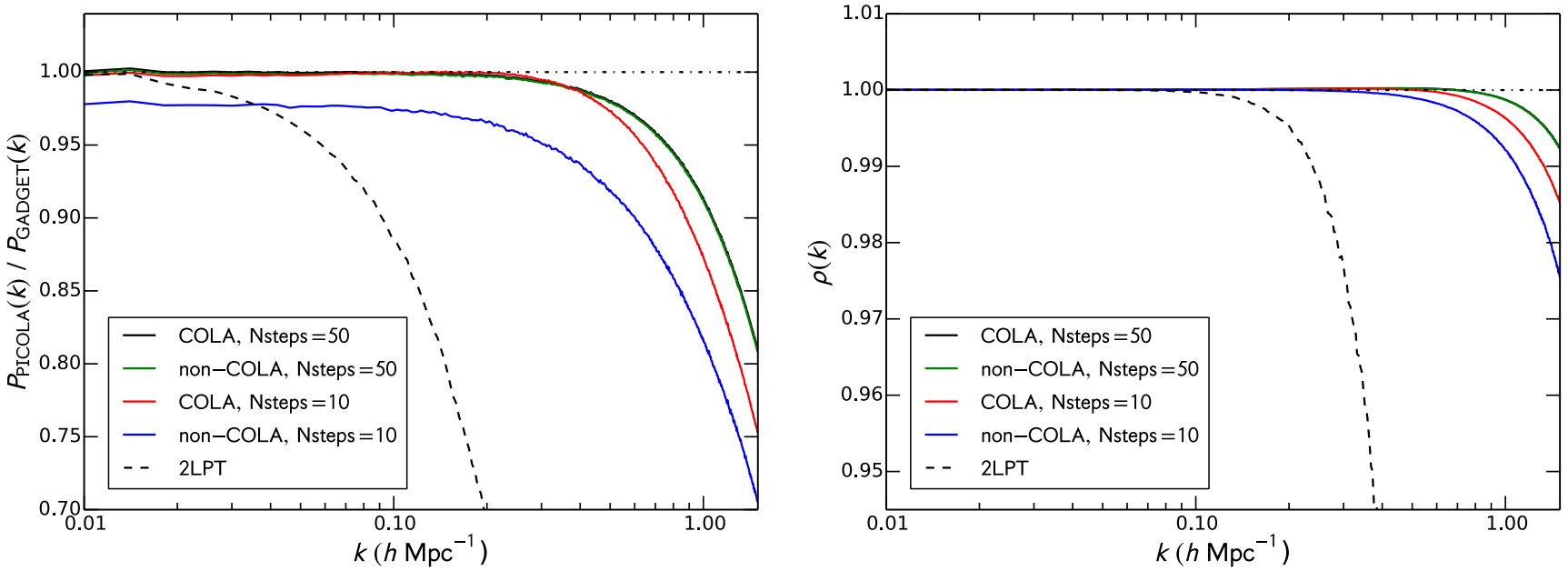


Figure 2.13: Plots of the power spectrum ratio and cross correlation between approximate realisations of the dark matter field, using the Particle-Mesh, 2LPT and COLA methods with 10 linear timesteps, and a Tree-PM realisation from GADGET-2.

caused by using a finite number of realizations). Figure 2.13 indicates that the real-space covariance matrix recovered from L-PICOLA is extremely accurate on all scales of interest to BAO and RSD measurements. Even where the cross-correlation between the L-PICOLA and GADGET-2 simulations deviates from 1, it still remains very high, such that the covariance matrix recovered from L-PICOLA would match extremely well that from a full ensemble of N-Body realisations even up to $k = 1.0\,h\,\mathrm{Mpc}^{-1}$.

In the same number of timesteps the Particle-Mesh algorithm cannot match the accuracy of COLA on any scale. Even on large scales there is a discrepancy between the PM and GADGET-2 runs, as there are not enough timesteps for the PM algorithm to fully recover the linear growth factor. This validates the reasoning behind the COLA method as the 2LPT solution provides the solution on linear scales but performs much worse than the PM algorithm on smaller scales. The time taken for a single timestep under both the COLA and PM methods is identical and as such the COLA method gives much better results for a fixed computational time.

Interestingly, however, the COLA and the standard PM algorithm converge if a suitable number of timesteps is used (50 in this case). When this many timesteps are used the PM code can accurately recover the linear growth factor and the non-linear clustering is greatly improved. Using a larger number of timesteps for the COLA run only affects the non-linear scales as the linear and quasi-linear scales are already fully captured. Using larger and larger numbers of timesteps has a diminishing effect on both algorithms, as the small scale accuracy becomes bounded by the lack of force resolution below the mesh scale. As the COLA method is already quite accurate for a few timesteps, increasing the number of timesteps for a fixed mesh size does not add as much accuracy as for the PM method alone. Incorporating the COLA mechanism into a Tree-PM code would negate this effect and it could be expected that increasing the number of timesteps used would then continue to increase the small scale accuracy beyond that achieved using the PM algorithm only.

Figure 2.14 compares the real and redshift-space cross correlation for the COLA and PM runs using 10 timesteps. The additional displacement each particle receives due to Redshift Space Distortions, s_{los}, is evaluated using

$$s_{los} = \frac{v_{los}}{H(a)a}, \tag{2.36}$$

where v_{los} is the line of sight velocity of each particle for an observer situated in the centre of the simulation box.

In all cases the accuracy of the simulation in redshift-space is worse than in real space. The 98% cross correlation continues only up to $k = 0.4\,h\,\mathrm{Mpc}^{-1}$. However, this is to be expected as, in addition to slightly under-predicting the spatial clustering of the dark matter particles, the approximate methods do not recover the full non-linear

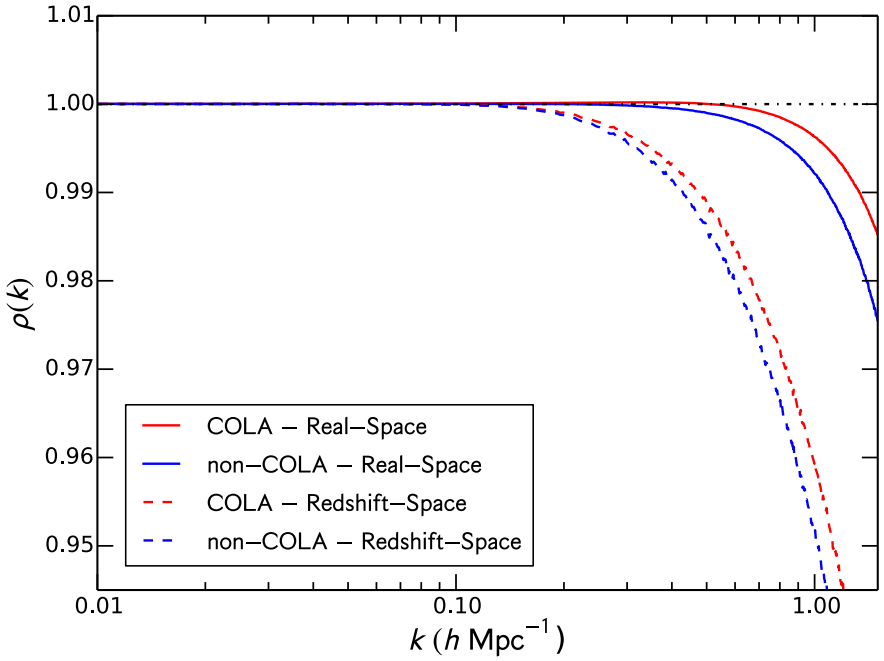


Figure 2.14: A comparison of the real- and redshift-space cross-correlations between approximate realisations of the dark matter field, using the Particle-Mesh and COLA methods with 10 linear timesteps, and a Tree-PM realisation from GADGET-2.

evolution of the particle's velocities as a function of time. The agreement in redshift space between the COLA method and the GADGET-2 run is still very good on all scales of interest to BAO and RSD measurements and the COLA method still outperforms the PM algorithm. Similarly we would expect the redshift-space covariance matrix to remain extremely accurate on these scales of interest.

2.6.2 Three-point Clustering

A further test is of the accuracy with which L-PICOLA recovers the three-point clustering of the dark matter field. In particular the reduced bispectrum,

$$Q(k_1, k_2, k_3) = \frac{B(k_1, k_2, k_3)}{P(k_1)P(k_2) + P(k_2)P(k_3) + P(k_3)P(k_1)}, \tag{2.37}$$

is used to quanify this, where $B(k_1, k_2, k_3)$ is the bispectrum for the periodic, cubic simulation.

In order to explore the agreement between GADGET-2 and L-PICOLA across a wide range of bispectrum configurations the reduced bispectrum ratio for L-PICOLA and GADGET-2 is plotted as a function of the ratios k_3/k_1 and k_2/k_1 for a variety of different values of k_1. This is shown in Figure 2.15. For clarity in this figure and to avoid double plotting the same configurations the conditions $k_1 \geq k_2 \geq k_3$ and $k_1 + k_2 + k_3 = 0$ are enforced.

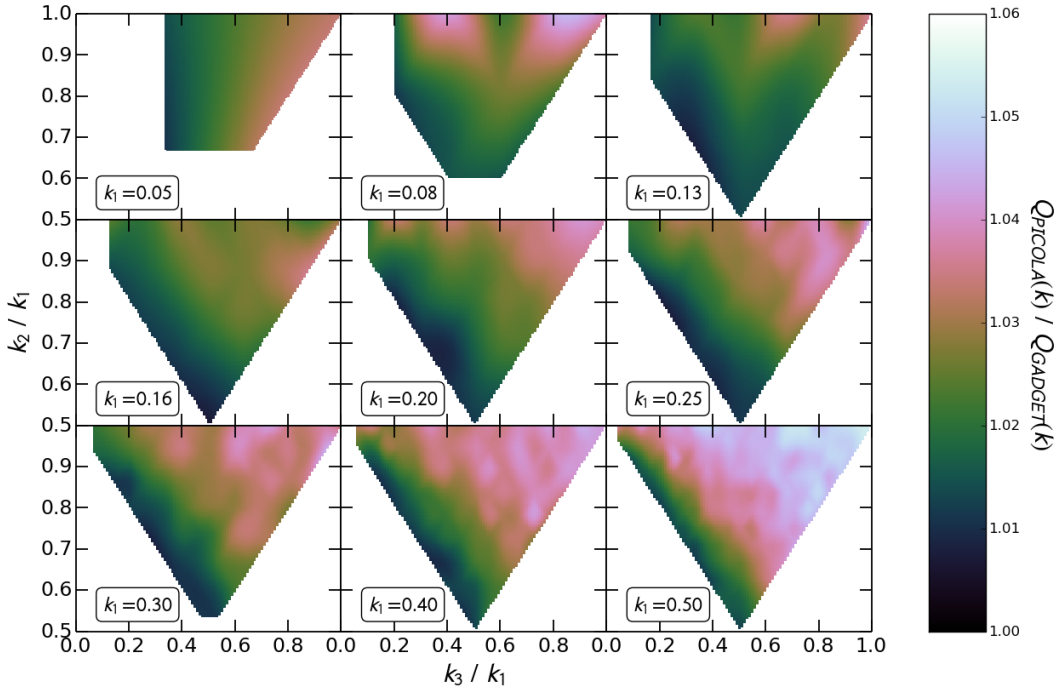


Figure 2.15: The ratio of the reduced bispectrum measured from the L-PICOLA and GADGET-2 simulations. The L-PICOLA runs use the COLA method with 10 linearly spaced timesteps. The bispectrum is plotted as a function of the ratios k_3/k_1 and k_2/k_1 for a range of k_1 values. This allows for the exploration of a wide range of triangle configurations. For reference, the top-left, top-right and bottom vertices of each plot correspond to squeezed, equilateral and folded configurations, whilst the left, and right and top edges correspond to elongated and isoceles triangles respectively.

From Figure 2.15 it can be seen that L-PICOLA is able to reproduce the reduced bispectrum to within 6% for *any* bispectrum configuration up to $k_1 = k_2 = k_3 = 0.5\,h\,\mathrm{Mpc}^{-1}$. This figure can also be used to identify the configurations that L-PICOLA reproduces with greatest and least accuracy. Regardless of the scale, the bispectrum in the squeezed, elongated or folded limit is reproduced extremely well, to within 2% on all scales. This is because these configurations contain large contributions from triangles with one or two large scale modes, which one can expect L-PICOLA to reproduce exactly. The least accurate regime is the equilateral configuration, with accuracy decreasing for smaller scales (larger k_1). This is because these triangles contain the biggest contribution from small scale modes in the simulation, which are not reproduced quite as accurately in L-PICOLA.

2.6.3 Timestepping and Mesh Choices

It should be noted that the convergence time of COLA depends intimately on the choice of timestepping and mesh size used and the accuracy after a given number of timesteps can vary based on the exact choices made. The representative run in Figure 2.13 uses the modified COLA timestepping, the value of $nLPT = -2.5$ suggested by Tassev et al. (2013) and a number of mesh cells equal to the number of particles.

Figure 2.16 shows how the accuracy of COLA is reduced when lower force resolution (less mesh cells) is used. The cases looked at are where the number of mesh points is equal to 1, 1/2 and 1/4 times the number of particles. A number of mesh cells larger than the number of particles is not considered as, from the Nyquist-Shannon Sampling Theorem, one should not expect any improvement in the clustering at early times, when the particle distribution is approximately grid based. Furthermore Peebles et al. (1989) and Splinter et al. (1998) advocate that there is little justification in using a force resolution higher than the mean particle separation due to the inevitable differences in clustering between different simulations caused by using a finite number of particles. For most practical applications of L-PICOLA it also becomes computationally infeasible to use a number of mesh cells much larger than the number of particles, due to the large increase in computational time for the Fourier transforms.

As expected, there is a reduction in the non-linear clustering accuracy as each mesh cell becomes larger, corresponding to a larger force smoothing. The large scales are still well recovered for all mesh sizes tested. Using smaller mesh sizes results in faster simulations and so for a given application of L-PICOLA a balance between mesh size and speed should be carefully considered based on the accuracy required and at which scales.

Figure 2.17 shows the effect of using timesteps linearly and logarithmically spaced in a and also the effect of using the modified timestepping (with $nLPT = -2.5$ still) compared to the standard Quinn et al. (1997) method.

In all cases the COLA method still outperforms the standard Particle-Mesh algorithm, although to differing degrees. In the case of identical timestepping choices between the COLA and PM runs the large scale and quasi-linear power is recovered much better. One point of interest is that using linearly spaced timesteps in the PM method reduces the accuracy on large scales below that of the logarithmically-spaced PM run, but greatly improves the non-linear accuracy, beyond even that of COLA with logarithmic steps. This is because using timesteps logarithmically spaced in a means the code takes more timesteps at higher redshift, where the evolution of the dark matter field is more linear. This means that the PM algorithm recovers the linear growth factor more accurately. Using linear timesteps results in more 'time' spent at low redshifts, where the evolution is non-linear and so the non-linear growth is captured more accurately, at the expense of

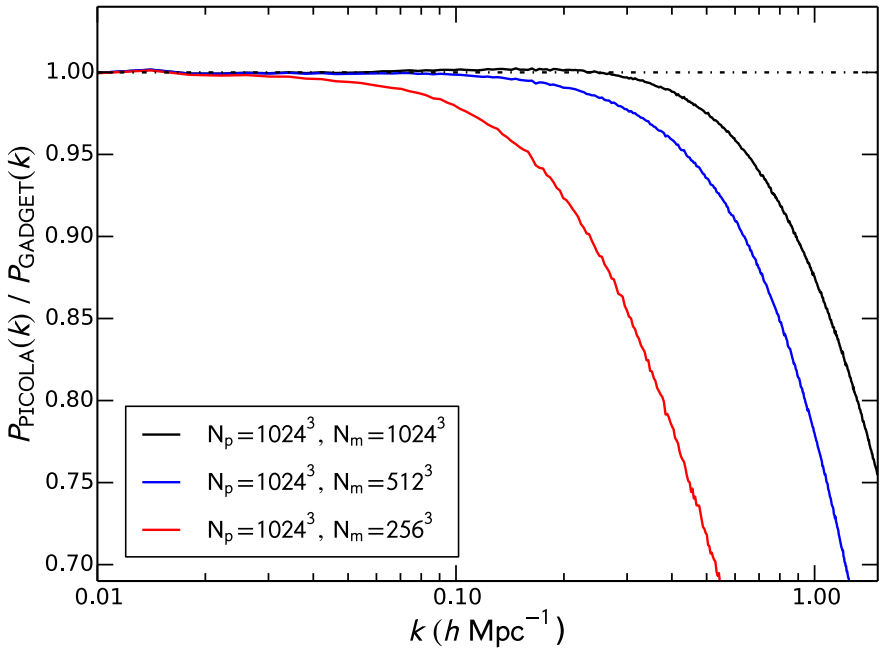


Figure 2.16: The power spectrum ratio between L-PICOLA dark matter fields using the COLA method and an N-Body realisation for different mesh to particle ratios. In all cases the simulation is run for 10 timesteps using linearly spaced COLA timesteps.

the large scale clustering. As the COLA method gets the large scale clustering correct very quickly, using linear timesteps to increase the non-linear accuracy is much more beneficial. Indeed, even more improvement is found using the modified timestepping method, Eq. 2.22, which emphasises the non-linear modes and corroborates the claims of Tassev et al. (2013).

It should be noted, however, that using the modified timestepping value puts additional emphasis on different growing modes, based on the value of $nLPT$, which can change the shape of the power spectrum. This is shown in Figure 2.18 where the power spectrum ratio between the N-Body and L-PICOLA runs for different values of $nLPT$ is plotted. Different values excite different combinations of decaying and growing modes, which are dominant at different cosmological times. This is shown in Figure 2.18 by the fact that different values produce slightly different power spectra. However, the cross-correlations for these runs are all very similar, indicating that the difference is non-stochastic and cannot vary from realisation to realization.

As such, though the 'correct' choice depends on the exact scales and statistics one wishes to reproduce with the mock realisations, this is not very important. The results can be calibrated afterwards simply by comparing two different simulations with different values of $nLPT$. For the fiducial case study in this chapter, a value $nLPT = -2.5$ shows

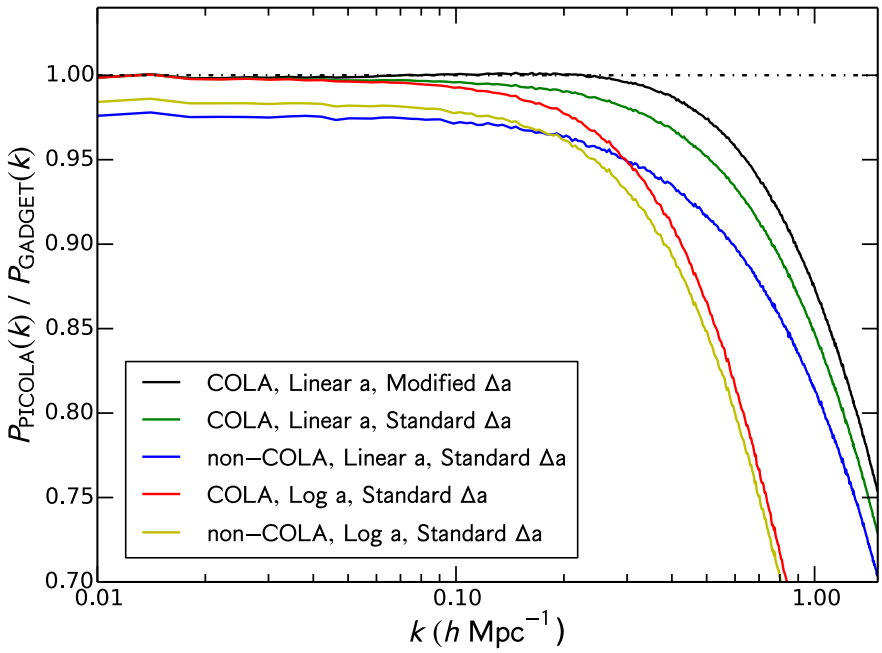


Figure 2.17: The power spectrum ratio between L-PICOLA dark matter fields and an N-Body realisation for different timestepping choices. This figure looks at the effect of using timesteps linearly and logarithmically spaced in a and using the modified method of Tassev et al. (2013) in place of the standard Quinn et al. (1997) timestepping. Also shown are a comparison of the COLA runs to standard PM runs using linearly and logarithmically spaced timesteps. In all cases the simulation is run for 10 timesteps.

reasonable behaviour on all scales.

Throughout this section it has been shown that the dark matter clustering recovered by L-PICOLA is extremely accurate on all scales of interest to BAO and RSD measurements. It is important to note however that when producing mock catalogues it is a representative galaxy field that is needed. In order to produce these L-PICOLA can be combined with other codes for identifying halos and populating the dark matter field with galaxies. Chapter 3 shows how the Friends-of-Friends algorithm (Davis et al., 1985) and Halo Occupation Distribution model (Berlind & Weinberg, 2002) were used to generate mock catalogues from L-PICOLA fields for the analysis of a low redshift galaxy sample. In this case no modification of the Friends-of-Friends linking length or the HOD model was needed. Other methods such as those presented by de la Torre et al. (2013), Angulo et al. (2014) or Kitaura et al. (2014) could also be used.

2.7 L-PICOLA Speed

As shown in the previous section, the COLA method outperforms both the 2LPT and Particle-Mesh algorithms in terms of the accuracy with which it reproduces the 'true' clustering recovered from a Tree-PM N-Body simulations. This section in turn highlights the transformation of the COLA method into a viable code for use with current and next generation large scale structure surveys by demonstrating the speed of L-PICOLA and showing how long it takes to produce a dark matter realisation compared to 2LPT.

This is done using a series of simulations with differing numbers of particles, box sizes and numbers of processors and by looking at the time taken in both the strong and weak scaling regimes. Strong scaling is defined as the change in the runtime of the code for different numbers of processors for a fixed simulation size, whereas weak scaling is the change in runtime for a fixed simulation size *per processor*. For the strong scaling test the same simulation specifications as for the accuracy tests are used, with numbers of processors equal to $\{8, 16, 32, 64, 128, 256\}$. The simulations used for the weak scaling are similar to those used for the strong scaling with additional details listed in Table 2.1. In all cases the number of mesh cells is fixed to the number of particles.[4]

All the run times are shown in Figure 2.19, for both the strong and weak scaling. In both cases the CPU time has been plotted in such a way that perfect scaling will result in a constant horizontal line (total CPU time summed across all processors for strong scaling and CPU time per processor for weak scaling). The top panel of this Figure shows the full CPU time taken for each run. The L-PICOLA runs generally take about 3 times longer

[4]All runs were performed on Intel Ivy Bridge CPU's on the SCIAMA high-performance computing cluster at the University of Portsmouth. More information can be found at http://www.sciama.icg.port.ac.uk/

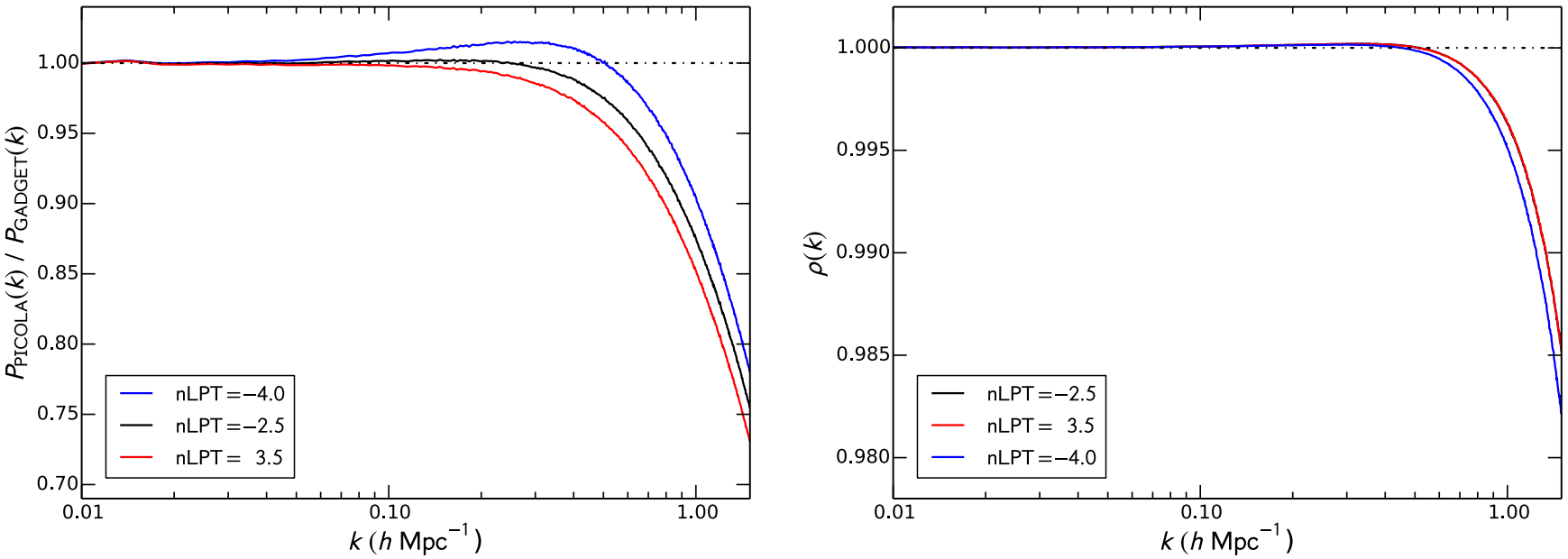


Figure 2.18: The power spectrum ratio and cross-correlation between approximate dark matter fields made with L-PICOLA and a GADGET-2 realisation for different values of $nLPT$ within the modifed COLA timestepping method. In all cases the simulations is run for 10 timesteps, using linearly spaced timesteps. The cross-correlation for both the $nLPT = -2.5$ and $nLPT = 3.5$ runs is similar enough to be indistinguishable.

Table 2.1: The specifications of the L-PICOLA and 2LPT runs used in the weak scaling tests. In all cases the number of mesh cells is fixed to the number of particles. All other simulation parameters are as used for the strong scaling runs and accuracy tests of Section 2.6.

N_{cpu}	$N_{particles}$	L_{box} ($h^{-1}\,\mathrm{Mpc}$)
2	256^3	192
4	322^3	246
8	406^3	304
16	512^3	384
32	644^3	484
64	812^3	610
128	1024^3	768

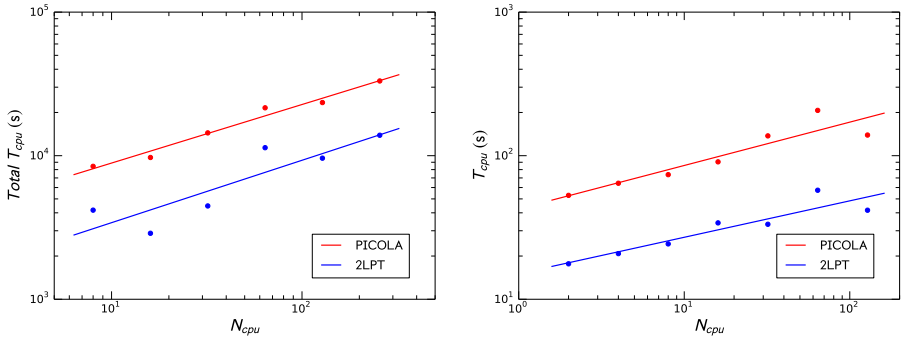
to run than a simple 2LPT realisation, however this is a relatively small cost compared to the difference in the accuracy of the methods.

In terms of the actual scaling, although L-PICOLA does not scale perfectly in either the strong or weak regimes, the increase in runtime with number of processors is still reasonable. A simple least squares fitting method is used to fit a linear trend to the CPU time as a function of the number of processors. From this gradients of 0.41 and 0.30 are found, compared to the ideal value of 0, for L-PICOLA in the strong and weak scaling regimes respectively. This trend can be extrapolated well beyond the fitting range., i.e., for a 2048^3 particle simulation in a $(1536\,h^{-1}\,\mathrm{Mpc})^3$ box run on 1024 processors the 'true' CPU time taken per processor is 348 seconds, which matches very well the predicted value of 345 seconds.

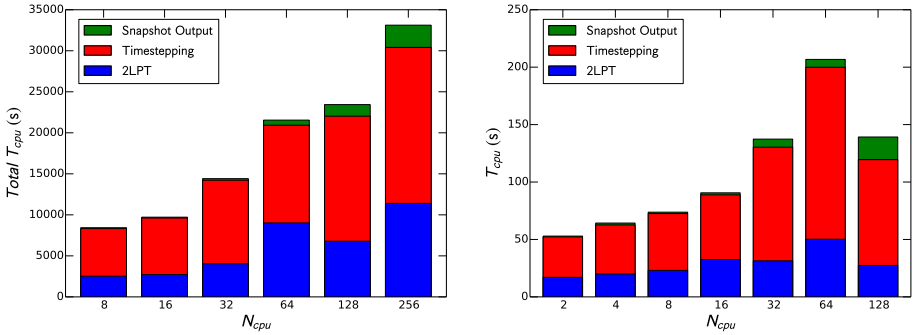
There exists some scatter in the runtimes for the simulations. This generally stems from the Fourier transforms involved, the efficiency of which depends on the way the mesh is partitioned across the processors. In the case where a number of mesh cells (in the x-direction) is used that is not a multiple of the number of processors (i.e., the $N_{cpu} = 64$, strong-scaling run) the time taken for the calculation of the 2LPT displacements and the interparticle forces during timestepping is increased.

2.7.1 Contributions to the Runtime

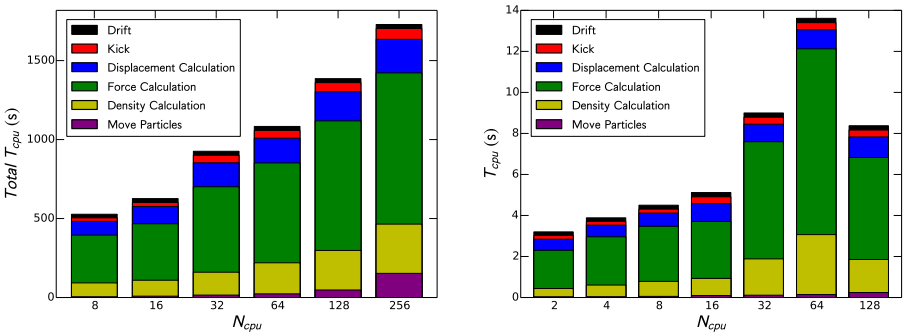
This is investigated further in the second and third rows of Figure 2.19. This figure shows the time taken for different contributions to the full run and to each timestep therein. This highlights the fact that the scatter occurs mainly during the 2LPT and force calculation



(a) Full Simulation



(b) Simulation Contributions



(c) Timestep Contributions

Figure 2.19: Plots showing the scaling of L-PICOLA in the strong (left) and weak (right) scaling regimes. For the strong scaling the total CPU time, summed across all processors, is plotted, whilst for the weak scaling it is the CPU time per processor. This means that ideal scaling would be shown as a constant horizontal trend as a function of the number of processors. Different panels show the total time taken for L-PICOLA compared to 2LPT simulations; the different contributions to the L-PICOLA runtime; and the contributions to a single L-PICOLA timestep.

parts of L-PICOLA as expected if it is due to the Fourier transform efficiency. Additionally this also suggests that the non-optimal strong and weak scaling does not stem from any particular part of the code, but rather due to the extra MPI communications needed when larger numbers of processors are used.

Looking at the contributions from the 2LPT, Timestepping and Output stages, there is some evolution with processor number in the 2LPT stage from the fact that the Fourier transforms require extra communications between different processors to transform the full mesh. There is also an increasing contribution from the Output stage of the code as larger numbers of processors are used. This is because of an option in the code to limit the number of processors outputting at once, stopping all processors outputting simultaneously. For these runs this has been set to 32 processors and as expected there is an increase in the time taken to output the data once the simulations use more than 32 processors due to the need for some processors to wait before they can output.

2.7.2 Contributions to a Single Timestep

Looking at the contributions to an individual timestep, the Drift, Kick and Displacement parts of the code are reasonably constant when the number of particles per processor remains constant. These consist mainly of loops over each particle and so this is to be expected. The Density Calculation and Force Calculation steps contain the Fourier transforms required for each timestep and as such are the biggest contributions to the time taken for a timestep. Looking at the strong scaling case there is an increase in both of these as a function of the number of processors, which indicates they are dominated by the MPI communications.

For the weak scaling there is also a large jump in the CPU time for both of these after 16 processors. This is another indication that the MPI communications are the cause of the scaling trends seen, as the architecture of the High Performance Computer we used for these tests is such that 16 processors are located on a single node and intra-node communication is much faster than inter-node. Once the code starts to require inter-node communication to compute the Fourier transforms, the CPU time increases.

Finally these scaling tests show that the Move Particles section of the code does not contribute much to the total time for each timestep, except where the number of inter-processor communications becomes large. This is due to the effort taken to produce a fast algorithm to pass the particles, whereas a simpler algorithm would result in a larger amount of time and memory needed to identify and store the particles that need transferring.

2.8 L-PICOLA Memory Consumption

Considerable effort has been made to reduce the memory footprint of L-PICOLA as much as possible, including the introduction of a compilation option to conserve as much memory as possible. When this option is used the memory consumption for a L-PICOLA run is reduced significantly and the mean memory per processor can be calculated reasonably simply.

With the optimal memory settings floating point precision is used for the particles and double precision for the mesh. The information for each particle consists of x, y and z coordinates, velocities in those same directions, and the ZA and 2LPT displacements in those directions, resulting in $M_p = 48\mathrm{Bytes}$ per particle. The main contributions to the memory arise from the particles and the mesh and the key parameters are the number of mesh cells, N_m, number of particles, N_p and the amount of buffer memory allocated to each processor to account for the non-uniformity of the particle distribution over processors at late times, b.

The code can be split into six distinct sections: the calculation of the initial 2LPT potentials; the calculation of the initial 2LPT displacements; the initialisation of the particles; the moving of particles across processors each timestep; the evaluation of the inter-particle mesh-based force each timestep; and the calculation of the particle displacements for each timestep. The corresponding memory requirements are:

$$M_{2LPT} = \frac{72N_m^2(N_m+2)}{N_{proc}} + 72N_m(N_m+2) + 4N_m^2 \tag{2.38}$$

$$M_{DISP} = \frac{48N_m^2(N_m+2) + 24N_p^3}{N_{proc}} + 48N_m(N_m+2) \tag{2.39}$$

$$M_{INIT} = \frac{(24+bM_p)N_p^3}{N_{proc}} \tag{2.40}$$

$$M_{MOVE} = \frac{(b+2(b-1))M_pN_p^3}{N_{proc}} \tag{2.41}$$

$$M_{DENS} = \frac{32N_m^2(N_m+2) + bM_pN_p^3}{N_{proc}} + 40N_m(N_m+2) \tag{2.42}$$

$$M_{NBODY} = \frac{(12+bM_p)N_p^3 + 24N_m^2(N_m+2)}{N_{proc}} + 24N_m(N_m+2) \tag{2.43}$$

The maximum memory required for an L-PICOLA simulation is the largest of these 6 contributions. A utility for calculating the memory requirements, even when using sub-optimal (in terms of memory) compilation options is provided with the public release of the code.

As an example Figure 2.20 shows the memory requirements as a function of number of processors for the L-PICOLA simulations used in Section 2.6. If there is 4GB of memory available per processor, this simulation can be run using only 25 processors if the

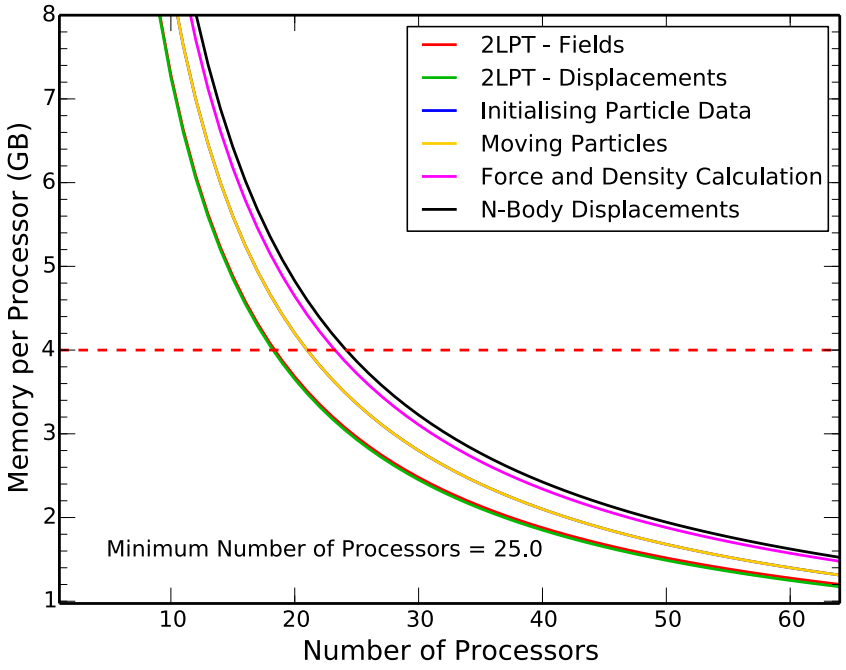


Figure 2.20: The memory requirements for the L-PICOLA run detailed in Section 2.6. The solid lines show the contributions from different sections of the code for varying numbers of processors, whilst the intersection of the dashed line with the solid lines gives the minimum number of processors required to run the simulation assuming there is 4GB of memory available per processor.

optimal compilation options are used (32 were used for the simulations in this chapter).

2.9 Summary

The main focus of this chapter has been the introduction and testing of a new N-Body code for simulating the evolution of dark matter. This code is designed primarily for generating large numbers of accurate dark matter matter simulations for use in estimating the statistical and systematic errors in large scale structure surveys. The historical context and motivation behind the creation of this code was presented in Section 2.1. In this section a brief review of existing methods for the fast generation of dark matter realisations was given, with emphasis on the algorithms that have been incorporated into the new code.

This code, L-PICOLA, is a memory conservative, planar parallelisation of the 2LPT and COLA algorithms which allow for fast yet accurate simulations of dark matter fields. The speed of this code is enabled by parallel algorithms for Cloud-in-Cell interpolation, Fast Fourier Transforms, and fast movement of particles between processors after each timestep. Additional features have also been included, such as the fast creation of initial conditions for other simulation codes, with optional primordial non-Gaussianity, and the ability to produce lightcone simulations, with optional replication of the simulation volume at run-time. These will be of particular use to future large scale structure surveys such as the Euclid survey (Laureijs et al., 2011).

In Section 2.5 the accuracy of the method L-PICOLA uses to produce lightcone simulations was tested and it was verified that the code's accuracy is not unduly affected by the approximations made to ensure a fast algorithm. The effect of replication on both the power spectrum and covariance matrix has also been tested using a set of 500 individual lightcone realisations. It is found that, due to the fact the replication procedure modifies the simulation volume without adding additional information, the power spectrum can suffer from ringing on the scale of the unreplicated box size and that the covariance matrix demonstrates the volume dependence of the unreplicated box size as opposed to the replicated volume. Simple corrections for both of these effects are presented and it is hypothesised that this is only a problem when analysing regions of the simulation larger than the unreplicated box size.

Section 2.6 compared the accuracy of L-PICOLA to the approximate 2LPT and PM methods and to a fully non-linear, Tree-PM GADGET-2 simulation. It was found that L-PICOLA performs much better than the 2LPT and PM algorithms, and that the power spectra from L-PICOLA agree with that from the GADGET-2 simulations to within 2% on all scales of interest to BAO and RSD measurements and to within 20% up to $k = 1.0\,h\,\mathrm{Mpc}^{-1}$. The reduced bispectrum from L-PICOLA also shows remarkable agreement with the GADGET-2 simulation, to within 6% for all configurations up to $k_1 = k_2 = k_3 =$

$0.5\,h\,\mathrm{Mpc}^{-1}$. It is found however that this agreement has some dependence on the exact type of timestepping used in the code.

The speed of L-PICOLA has also been compared to 2LPT. The remarkable accuracy of L-PICOLA comes at only a small cost to speed compared to 2LPT. L-PICOLA exhibits reasonable scaling properties in the strong and weak scaling regimes, even up to large numbers of processors. These trends are dominated by the need for extra inter-processor communication when using large numbers of processors. The speed of L-PICOLA compared to a fully non-linear GADGET-2 simulation is given in the next chapter (Section 3.2.4), as is a comparison of the halo catalogues recovered from both codes (Section 3.3).

Chapter 3

Producing Mock Catalogues for the SDSS Main Galaxy Sample

Measurements of the Cosmic Microwave Background have allowed for tight constraints on the nature of the universe and its fundamental components close to its inception (Planck Collaboration et al., 2014b). However, current, plausible models that are consistent with the high redshift results of the Planck Collaboration et al. (2014b) may diverge greatly at $z = 0$. In order to understand the observed acceleration of the cosmic expansion rate (see Riess et al. 1998; Perlmutter et al. 1999 for early detections and Weinberg et al. 2012 for a review of observational probes), and the growth of structure throughout cosmic time, accurate measurements are needed across a range of redshifts. In particular, low-redshift measurements offer the most promising route to determining the nature of the late time acceleration.

Motivated by this, the work in the following two chapters details the use of a low-redshift sample of galaxies with measured redshifts from which the BAO and RSD signals can be modelled and measurements of the expansion rate and growth rate of structure obtained. By applying the latest modelling and analysis techniques these measurements complete the set of accurate BAO-distance and RSD constraints that can be made from current data. This chapter justifies the choice of data and presents the production of a set of mock catalogues used to aid the analysis. These mock catalogues enable testing of the models and fitting procedure used to recover the BAO and RSD measurements, and are used for systematic tests on the measurements. This in turn allows the robustness of the measurements to be quantified. The measurements themselves are presented in the next chapter.

The first section in this chapter presents the data used for this application. Subsequent sections cover the exact procedures for producing the dark matter fields with the L-PICOLA code from the previous chapter, converting the dark matter fields to halos, pop-

ulating these halos with galaxies and subsampling the galaxies to match the angular and redshift distribution of the data. This chapter then provides comparisons of the clustering of the data and mock catalogues and quantifies the accuracy with which the mocks are able to reproduce the clustering of the data. Finally, systematic tests will be performed using the mocks.

3.1 The Sloan Digital Sky Survey Data Release 7 Main Galaxy Sample

The Sloan Digital Sky Survey (SDSS; York et al. 2000) Data Release 7 (DR7; Abazajian et al. 2009) contains the completed data set of the full SDSS-I and SDSS-II surveys. These surveys obtained wide-field CCD photometry (Gunn et al. 1998, 2006) in five passbands (u, g, r, i, z; Fukugita et al. 1996), internally calibrated using the 'uber-calibration' process described in Padmanabhan et al. (2008). This process resulted in photometric imaging of 357 million unique objects over a total footprint of 11,663 deg^2. From this imaging data, galaxies within a footprint of 9380 deg^2 (Abazajian et al., 2009) were selected for spectroscopic follow-up as part of the Main Galaxy Sample (MGS; Strauss et al. 2002). This in turn resulted in a flux-limited low-redshift sample of $\sim 930,000$ galaxies, which, to good approximation, consists of all galaxies with $r_{\rm pet} < 17.77$, where $r_{\rm pet}$ is the extinction-corrected r-band Petrosian magnitude. A further extension to this program, spectroscopically measuring Luminous Red Galaxies at higher redshift, was carried out with selection as reported in Eisenstein et al. (2001), although this sample is not used for the work in this thesis.

The SDSS DR7 MGS data used was extracted from the DR7 value-added galaxy catalogues hosted by New York University[1] (NYU-VAGC). These catalogs were created following the methods described in Blanton et al. (2005). They include K-corrected absolute magnitudes, determined using the methods of Blanton et al. (2003), and detailed information on the angular selection function. The galaxy sample used for the cosmological anaylses in this thesis were selected from the NYU-VAGC 'safe0' catalogue. The 'safe0' catalogue only uses galaxies with $14.5 < r_{\rm pet} < 17.6$. The $r_{\rm pet} > 14.5$ limit ensures that only galaxies with reliable SDSS photometry are used and the $r_{\rm pet} < 17.6$ allows a homogeneous selection over the full footprint of 7356 deg^2 (Blanton et al., 2005). The end result is a sample containing $\sim 600,000$ galaxies.

Further corrections within the 'safe0' catalogue account for fiber collisions, an observational effect where no two fibers on the same spectroscopic tile can be placed more closely than $55''$. If uncorrected this results in a diminished measurement of the small

[1] http://sdss.physics.nyu.edu/vagc/lss.html

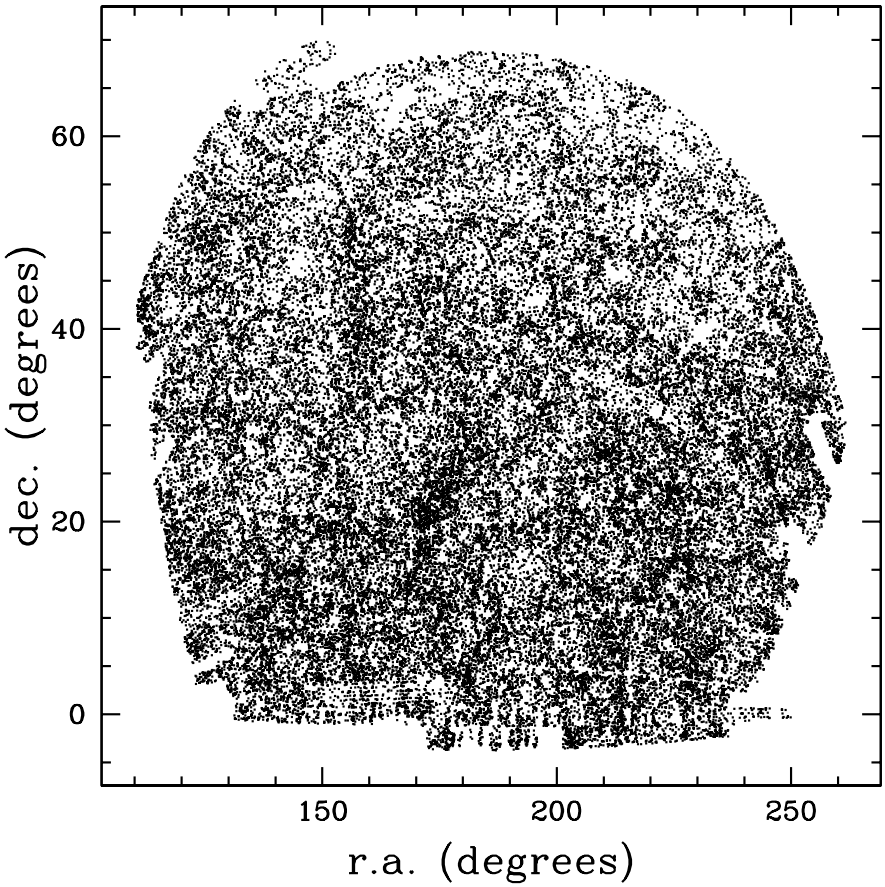


Figure 3.1: The right ascension and declination positions (J2000) of the 63,163 $z < 0.2$ SDSS DR7 MGS galaxies. Their footprint occupies 6813 deg^2. This plot is taken from Ross et al. (2015)

scale clustering of the galaxies. Galaxies that did not obtain a redshift due to fibre collisions are instead given the redshift of their nearest neighbour.

3.1.1 Creating the $z < 0.2$ MGS Galaxy Catalogue

The creation of the $z < 0.2$ MGS, which from hereon will simply be denoted MGS, is the result of several additional cuts to the NYU-VAGC 'safe0' catalogue. Only data from the contiguous area in the North Galactic cap and data occupying areas where the completeness is greater than 0.9 was used, to ensure a 'pure' sample with relatively simple window function. These cuts reduce the footprint to 6813 deg^2. The angular positions of the galaxy sample used are plotted in Fig. 3.1.

Further cuts based on colour, magnitude, and redshift were then made to balance the following motivations:

1. Create a sample that is at a lower redshift than is probed by the SDSS-III BOSS. Therefore only galaxies with $z < 0.2$ were used.

2. Be able to reliably simulate the clustering of the galaxies in the sample. This requires a reasonably constant galaxy density as a function of redshift, $n(z)$, and that galaxies occupy dark matter halos with masses $M_{\rm halo} > 10^{12}\,M_{\odot}$, which is the minimum halo mass that our simulations can reliably achieve.

3. Minimize the fractional uncertainty expected for measurements of $P(k)$, balanced against the above two concerns. This is achieved by maximizing the galaxy density across the survey

Balancing these motivations, the sample was defined to have $0.07 < z < 0.2$, $M_r < -21.2$ and $g - r > 0.8$, where M_r is the r-band absolute magnitude provided by the NYU-VAGC. The resulting sample contains 63,163 galaxies, with the large reduction in the number of galaxies resulting primarily from the r-band absolute magnitude cut. The luminosity and colour cuts make the sample more homogeneous as a function of redshift and increase the clustering amplitude of the sample. The increase in clustering amplitude implies an increase in the mass of the typical halo hosting the galaxies. The results presented in Zehavi et al. (2011) (e.g., their figure 10), suggest a negligible fraction of galaxies matching the selection criteria occupy halos with $M < 10^{12}\,M_{\odot}$ and therefore imply one should be able to reliably simulate the sample. The $z > 0.07$ limit is applied as, due to the $r_{\rm pet} > 14.5$ cut, the number density of the MGS drops sharply for $z < 0.07$.

The $n(z)$ of the MGS is displayed in Fig. 3.2. The $n(z)$ was fit to a model with two linear relationships and a transition redshift, given by

$$n(z) = 0.0014z + 0.00041; z < 0.17 \tag{3.1}$$

$$n(z) = 0.00286 - 0.0131z; z \geq 0.17, \tag{3.2}$$

which provides a good representation of the data, as the χ^2 is 25 for 22 degrees of freedom (26 $n(z)$ bins and four independent model parameters). This function is used to create the mock catalogues, assign redshift-dependent $w_{\rm FKP}$ weights for the clustering measurements and assign redshifts to the random catalogue used to measure the clustering of mock samples. This fit is not used to assign redshifts to the angular positions in the random catalogue used to measure the clustering of our data sample however. Instead redshifts are randomly selected from the galaxy catalogue, thus allowing for any further observation-dependent fluctuations. Ross et al. (2012) show that this imparts negligible bias on BOSS clustering measurements. It will be shown later that this also makes a negligible difference for the MGS. After cutting the sample, the effective redshift of the MGS is $z_{\rm eff} = 0.15$.

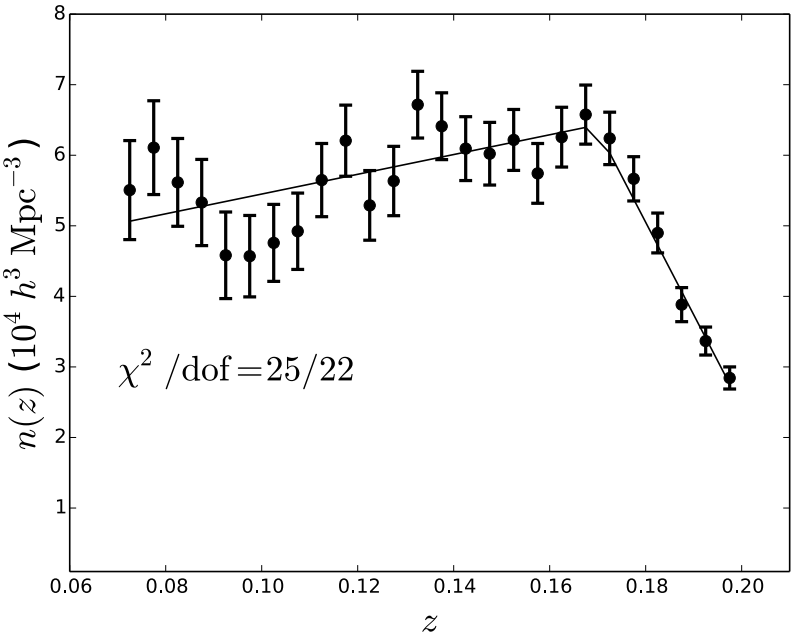


Figure 3.2: The number density of the MGS plotted as function of redshift. The error-bars represent the standard deviation of $n(z)$ for the mock catalogues described in this chapter. The curve is the best-fit model assuming two linear relationships with a transition redshift, which is a good fit to the data. This plot was taken from Ross et al. (2015)

3.2 Producing Dark Matter Fields

This section describes how L-PICOLA was used to produce dark matter fields for the MGS. As described in the previous chapter L-PICOLA can generate a dark matter field from given only a few input parameters. The following subsections describe and justify the choice of parameters used for the dark matter runs.

3.2.1 Fiducial Cosmology

As is commonly done when producing mock catalogues, the fiducial cosmology used to generate the dark matter fields for the MGS sample was chosen to match the cosmology used for modelling and analysing the data itself. When performing tests of fitting procedures on the mocks this then allows easy comparison between best-fit parameters returned from modelling and fitting the mocks and the expected values for those mocks. The fiducial cosmology chosen was that of a flat ΛCDM universe with $\Omega_m = 0.31$, $\Omega_b = 0.048$ and $H_0 = 67.0\,\mathrm{kms}^{-1}\mathrm{Mpc}^{-1}$ at redshift $z = 0.0$, $n_s = 0.96$, and with the clustering strength normalised at $z = 0$ using $\sigma_8 = 0.83$, where σ_8 denotes the matter variance within a spherical top-hat filter of radius $8\,h^{-1}\,\mathrm{Mpc}$. This cosmology was designed to coincide well with the best-fit cosmological parameters from the Planck 2014 results (Planck Collaboration et al., 2014b). Initial density perturbations for the MGS dark matter simulations were set up at an redshift $z = 9.0$ using a linear matter power spectrum generated using the Boltzmann code CAMB (Lewis et al., 2000). The low initial redshift of the simulations, compared to the commonly used values of $z \approx 50$ or $z \approx 100$, matches that used in the L-PICOLA accuracy tests of Section 2. Second-order Lagrangian Perturbation Theory provides a very accurate solution to the dark matter equation of motion at $z > 9.0$, when the evolution is mostly linear. Hence, in the COLA method, timestepping can start at much later times than in a standard N-Body simulation as only the non-linear evolution, most prevalent at low redshifts, needs to be solved.

3.2.2 Mass Resolution

Owing to the low redshift of the MGS sample, the maximum comoving distance from the observer to any galaxy is only $\sim 570\,h^{-1}\,\mathrm{Mpc}$. However, the angular extent of the MGS dataset is such that one axis of the simulation box must extend $\sim 570\,h^{-1}\,\mathrm{Mpc}$ *in each direction* to include all galaxies, requiring a total length of twice the maximum comoving distance. To minimize complications arising from a non-cubic box during the dark matter simulation, and avoid complex transformations of the survey geometry to fit in a box of intermediate length, it was decided to simulate a box large enough to cover the full-sky out to the maximum redshift of the MGS data. This also allowed replication of the MGS

survey within a single simulation, such that multiple mock catalogues can be obtained from a single dark matter realization. This reasoning resulted in a cubic simulation box of edge length $1280\,h^{-1}\,\mathrm{Mpc}$.

By design, the MGS sample is limited to include galaxies that are expected to reside within halos of mass, $M_{halo} > 10^{12} M_{\odot}$ based on the color and luminosity dependent Halo Occupation Distribution (HOD) fits to the clustering of SDSS galaxies by Zehavi et al. (2011). As such, the mass resolution of the MGS mock catalogues must be such that halos of this mass or greater will contain a large enough number of particles to allow for accurate identification of halos, minimizing the impact of mis-identifying clusters of particles randomly situated close together as dark matter halos. To meet this demand each dark matter field was simulated using 1536^3 particles, which results in a particle mass of $\sim 5 \times 10^{10}\,h^{-1}\,\mathrm{M}_{\odot}$.

3.2.3 Redshift Evolution

All the dark matter simulations are evolved using the COLA mechanism with $n_{LPT} = -2.5$, from the initial redshift up to the effective redshift of the MGS sample, $z = 0.15$, using 10 timesteps linearly spaced in the scale factor, a.

3.2.4 Comparison to GADGET-2 Simulations

Computing Time

Each simulation takes around 20 minutes (including halo-finding) on 256 cores. In terms of the actual computing time used, each L-PICOLA run took ~ 25 CPU-hours compared to the $\sim$ 27600 CPU-hours required for a comparison GADGET-2 run using the same initial conditions and simulation parameters. However, it should be noted that the actual (wall)time taken for the GADGET-2 run was not 1000 times that of a single L-PICOLA run, because the memory requirements of GADGET-2 are also larger than those of L-PICOLA, requiring more processors to run (384 in this case).

Two-point Clustering

Fig. 3.3 shows the power spectrum of the dark matter fields for one of the L-PICOLA simulations and for a Tree-PM N-Body simulation performed using GADGET-2 (Springel, 2005). Both simulations use the same initial conditions and the same mesh resolution. We can see that the power spectra agree to within 2 percent across all scales of interest to BAO and RSD measurements and the agreement continues to within 10 percent to $k \sim 0.8\,h\,\mathrm{Mpc}^{-1}$.

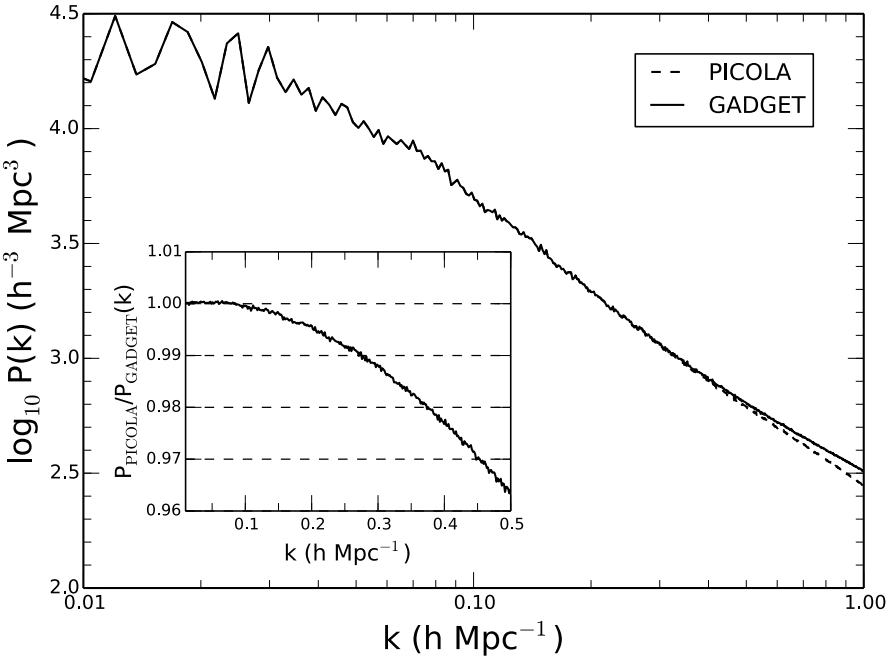


Figure 3.3: The power spectra of the dark matter field in a cubic box from the MGS L-PICOLA and GADGET-2 simulations described in the text. There is good agreement between the two even into the non-linear regime.

3.3 From Dark Matter to Halos

Identification of virialised dark matter halos in each simulation was carried out using the well-known Friends-of-Friends algorithm (FoF; Huchra & Geller 1982; Einasto et al. 1984; More et al. 2011). In the algorithm particles are joined if they have a separation less than $b\bar{l}$, where $\bar{l}$ is the mean particle separation and b is the 'linking length', a free parameter which must be specified for the particular simulation. This algorithm is a popular choice due to its simplicity; the resulting grouping of particles depends only on this single free parameter, although the exact value that should be chosen for this is a point of some contention. The most commonly used value is $b = 0.2$ (Frenk et al., 1988), which is adopted when creating the MGS mocks. Lacey & Cole (1994) rationalize this choice of linking length using the approximation that structures found with this linking length are bound by surfaces of constant density given by

$$\rho = \frac{3\bar{\rho}}{2\pi b^3} \tag{3.3}$$

where $\bar{\rho}$ is the mean density of the simulation. Assuming virialised regions have a density profile given by an isothermal sphere with $\rho \propto r^{-2}$ such that $\langle\rho\rangle = 3\rho$, one finds that the expected overdensity constrast between FoF-identified halos and the backfround density is $\langle\rho\rangle/\bar{\rho} = 180$ for $b = 0.2$. This coincides well with the value given by Peebles (1980)

for a virialised object resulting from spherical top-hat collapse in an Einstein-De Sitter universe, $\langle\rho\rangle/\bar{\rho} = 18\pi^2$.

Additionally, when using the FoF algorithm a decision has to be made at which point to define a grouping of particles as a halo. This takes the form of a minimum number of particles required within a bound structure for that structure to be called a halo. In practice the larger this value, the more confident one can be that the collection of particles truly represents a halo formed at a peak in the overdensity field, rather than a collection of particles randomly situated close together due to shot-noise in the simulation.The cost of using a high value is that, depending on the simulation resolution, only halos above a given mass are captured. One possibility is to relate this number of particles to some minimum overdensity threshold, however for the MGS mocks a threshold of 20 particles was applied. This is twice that used for mocks produced for the BOSS LOWZ galaxy sample Manera et al. (2015), and as such it is expected that the contribution of noise in the dark matter fields to the resultant halo catalogues is negligible.

Due to the nature of the algorithm, the FoF algorithm does not enforce sphericity on the halos it finds. Whilst this can be beneficial as it recreates the triaxiality expected in real dark mater halos, this can present a problem at later stages of mock catalogue production as the HOD model usually assumes the halos are spherical. Dark matter simulations also have small scale noise associated with their finite resolution and this in turn can result in cases where two halos which would be distinct in more detailed simulations are joined together by one or two particles that are only associated with the halos due to random noise in the dark matter field. This can result in 'dumbbell' shaped halos and loss of accuracy in the recovered halo mass function.

Accordingly, there are a a range of different methods for finding halos on top of the FoF method, based on finding overdensity peaks in the dark matter fields that satisfy some threshold (Lacey & Cole, 1994; Klypin & Holtzman, 1997), or using phase-space information to identify shell crossing (Falck et al., 2012). There are also numerous implementations of these three methods, however in the summary and comparison of these methods, Knebe et al. (2011) found that the statistical properties, in terms of the clustering and mass function, of the halos recovered by these distinct methods agree surprisingly well. Because of this the FoF algorithm was retained for the MGS mock catalogues in the interest of speed and simplicity. However, to deal with the volume of data required to produce the mocks in a timely manner a fast MPI-implementation of this algorithm, that complements the output from L-PICOLA, was developed.

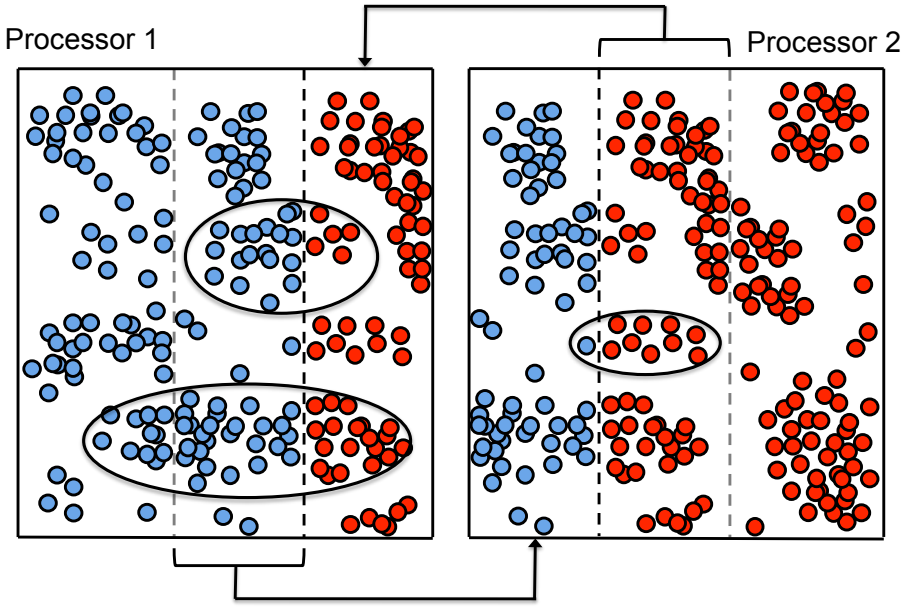


Figure 3.4: A pictoral representation of the sharing of particles across boundaries in CM_HALOFINDER, which is necessary to ensure that all constituent particles of a halo near the boundary are included in the FoF algorithm. Circled groups of particles represent halos containing contributions from both processors. Only the processor containing the centre of mass outputs the halos, and it is this copy of the halo that is circled.

3.3.1 CM_HALOFINDER

This section presents the FoF code developed to produce the MGS mock catalogues. As a complementary code to L-PICOLA this was designed to deal with all output formats of this latter code and to also analyse simulations on the past lightcone. The inner workings of the code can be split into two distinct parts.

Preparing the Dark Matter Field

Due to the nature of the FoF algorithm, parallelisation presents a problem. Ideally a simulation would be spatially split over many processors and each processor would identify the halos in its portion of the full simulation. With the FoF algorithm however this is not fully possible as particles near the boundaries of each processor will likely be part of the same halo as those on a neighbouring processor. This can be corrected by ensuring copies of particles within some length scale, h, of the processor boundary are given to all other neighbouring processors which share that boundary. A graphical representation of this is shown in Figure 3.4.

As long as this boundary scale h is larger than the maximum extent of any halo within the simulation, then it is guaranteed that at least one of the processors sharing a boundary

will have every particle associated with a halo near that boundary. Furthermore, it is logically true that one (and only one) of the processors will contain a complete version of that halo that has a centre of mass outside of the boundary region. To avoid duplicates, it is this processor that stores and outputs the halo. Halos of this kind have been highlighted in Figure 3.4, on the processor that would output them.

Whilst logically sound, this algorithm does present a problem in that the boundary scale must be larger than the largest halo, and must be present for every edge that is shared between processors. This could result in the case that multiple copies must be made of every particle in the simulation. This is turn could result in asking each processor to analyse more particles than can be stored in adjacent memory, making the algorithm unfeasible. This is especially true of the case when the simulation is split into slices, as in L-PICOLA, and the width of the slices approach the size of a single halo, which could often be the case for particular dense simulations or low numbers of processors.

To mitigate this, in CM_HALOFINDER the original dark matter simulation is not split into slices but rather into cubes. For a fixed number of processors and particles this allows similar levels of parallelism whilst retaining a reasonable ratio between the boundary size and the edge length of the simulation region on each processor, even when the simulation is dense. The cost of this is increased complexity when reading-in, organising and preparing the dark matter field as each processor now has 26 neighbours (and boundary regions) as opposed to only 2. In other words, for a cube cut into 3 slices along a single axis the central slice has two neighbouring slices, whereas for a cube cut into 3 along *each* axis the result is 27 sub-cubes, such that the central cube now has 26 neighbouring cubes.

Overall, the method for reading in and assigning the particles to processors proceeds as follows:

1. Decide which subset of processors will be reading in the files.
2. Loop over all the dark matter files. For each file:
   (a) Read in and store all the particles in the file.
   (b) Count the number of particles being sent to each processor. For each particle in the file:
      i. Calculate which processor each particle should go to. This is done by calculating the processor number, n, for each particle located at position x, y, z within the simulation, using

$$n = N_y N_x \left\lfloor \frac{N_z z}{L_z} \right\rfloor + N_x \left\lfloor \frac{N_y y}{L_y} \right\rfloor + \left\lfloor \frac{N_x x}{L_x} \right\rfloor \tag{3.4}$$

      where N_x, N_y, N_z and L_x, L_y, L_z are the number of processors and the length of the box in the x, y and z directions respectively. $\lfloor x \rfloor$ denotes

rounding down x to the nearest integer.

ii. Displace the x, y and z positions of each particle by +/- the boundary size (6 iterations) and recalculate the processor for the particle.

iii. If the displaced particle position is on a different processor to the main processor then the particle is within a boundary and needs duplicating. Generate a 3-element x,y,z vector for the particle where each element is -1,0 or 1 if the particle is within a boundary in the negative direction (-1), positive direction (1) or not at all.

iv. Increment the counter for the number of particles being sent to each processor based on the 3-element vector. If there are any non-zero elements in the vector then increment the counters for each non-zero element individually and then in combination. I.e., for a particle with vector (1,-1,-1), increment the counters as if the particle had vectors (0,0,0), (1,0,0), (0,-1,0), (0,0,-1), (1,-1,0), (1,0,-1), (0,-1,-1) and (1,-1,-1). This ensures that, as the processors hold cubic regions of space, the particle is shared with *all* neighbouring processors.

(c) Send the counters to every processor so that it knows how many particles to receive and from whom.

(d) Repeat the above procedure for each particle and this time store the particles (and duplicates) in an array such that the first set of particles will be sent to the first processor.

(e) Send the particles to each processor, which them stores them. If the boundaries of the simulation box are periodic we have to 'wrap' the particles to that their relative displacements are correct.

3. Check that all the expected particles have been read in and stored.

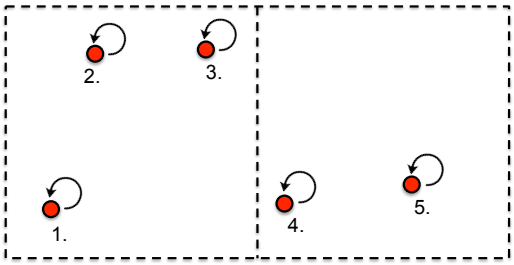
Once this procedure has been performed, all the processors will contain only particles corresponding to the portion of the simulation it has been assigned, plus those particles on neighbouring processors within the boundary region. From here the FoF algorithm can proceed on each processor independently.

Identifying Halos

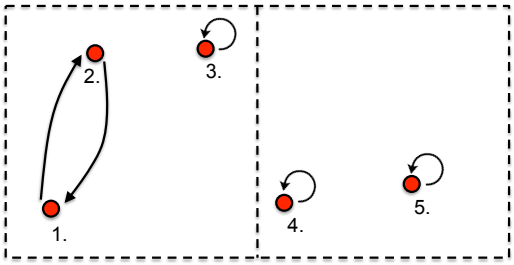
To identify the halos on each processor in an efficient manner, the particles are assigned to cells and a linked list is created for each cell so that each particle in that cell 'points' to the next one. The cell size is chosen based on the mean particle separation, such that on average each cell should contain a single particle. As the linking length is given by some

fraction of the mean particle separation, this in turn means that when searching for two particles within the linking length we only need to search the neighbouring cells. This is further simplified by the fact that if we start from a corner of the processor's spatial region, we only need to search the 'forward' cells on each processor (i.e. in the +x, +y, and +z directions). The particles in the cells 'behind' will, by design, have already been grouped together. For each processor, the algorithm relies on each particle pointing to a single other particle in the same halo, such that for a halo of N particles there are N pointers. The algorithm works as follows:

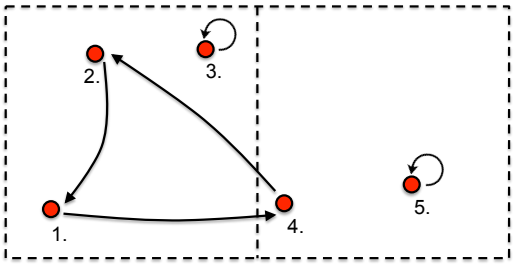
1. The starting point of the algorithm is that each particle points to itself. The algorithm then proceeds by looping over all cells within a processor.



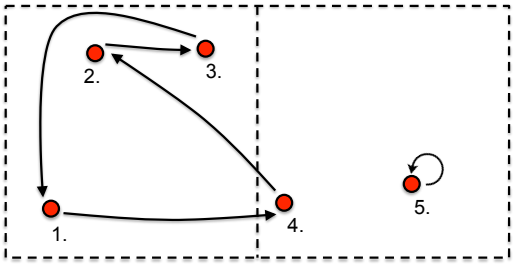
2. Look at the first particle in the first cell. If there is no particles in that cell then move on. Otherwise for all other particles in the same cell, check to see if this particle is within the linking length of the first particle. If so the pair exchange pointers, such that each particle is now pointing to where the other was pointing.



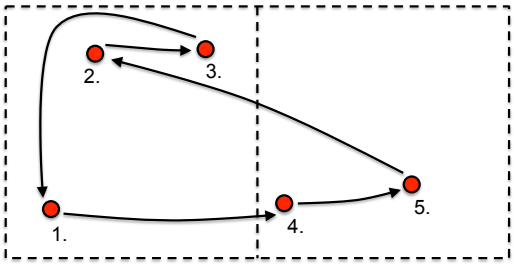
3. For all particles in the 'forward' cells, check to see if this particle is within the linking length of the first particle. If so the pair exchange pointers.



4. Look at the second particle in the cell, if one exists. This has already been compared to the first particle. For all other particles in the same cell (excluding the first), check to see if this particle is within the linking length of the *second* particle. If so then check to see that this has not already added to the halo based on proximity to the *first* particle and add if required. Do the same check for all the 'forward' cells. Then repeat this whole step for all other particles in the same cell.



5. Look at the next cell and repeat this procedure.



Overall, this process creates a series of closed pointer loops. Following these loops then gathers all the particles associated with a given halo. Summing the individual particle properties for each halo returns the centre of mass position, the velocity, the moment of inertia tensor and velocity dispersion tensor for that halo. At this point a check is also made that the boundary region used was large enough, by ensuring that no constituent particles of any halo with a centre of mass outside of the boundary region are within a linking length of the edge of the processor. If the boundary is not large enough then the simulation has to be rerun. In practice a value of $15\,h^{-1}\,\mathrm{Mpc}$ is often found to be enough. Whilst this may seem much larger than a typical halo, the tendency of the FoF algorithm to link two distinct halos together via a small number of particles, creating 'dumbbell' shaped halos, contributes significantly to this.

3.3.2 Application to BOSS-LOWZ Mock Catalogues

Originally, the CM_HALOFINDER code was developed to identify halos in 2LPT simulations which in turn were used to develop mock catalogues for the BOSS-LOWZ galaxy sample. The Baryon Oscillation Spectroscopic Survey (BOSS) is part of the third iteration of the Sloan Digital Sky Survey. Its goal is to provide 1.35 million spectroscopic measurements of galaxies imaged in previous incarnations of the SDSS. Following the

targetting algorithm of Eisenstein et al. (2001), BOSS targets two distinct samples of galaxies at different redshift. The first of these is the 'CMASS' sample which consists of galaxies with confirmed redshifts $0.4 \lesssim z \lesssim 0.7$. The LOWZ sample on the other hand contains galaxies with $0.25 \lesssim z \lesssim 0.4$. Data Release 11 (DR11) from the BOSS contained the first public release of the BOSS-LOWZ sample and cosmological measurements from this data, resulting from a 2% measurement of the BAO scale, is presented in Tojeiro et al. (2014).

In order to produce these measurements, Manera et al. (2015), closely following the procedure in Manera et al. (2013), created a set of mock galaxy catalogues from which the measurement covariance was estimated. 500 dark matter fields simulations were created at $z = 0.32$ using the PTHALOS algorithm Scoccimarro (1998). Halos were then identified in these simulations using CM_HALOFINDER, however careful calibration of the linking length was required to recover the halo mass function; because the PTHALOS method uses only 2LPT to approximate the dark matter clustering, the dark matter particles have not collapsed sufficiently to be gathered using $b_{N-Body} = 0.2$. Using the spherical collapse approximation, Manera et al. (2013) related the required linking length to the standard value via

$$b_{2lpt} = b_{N-Body} \left(\frac{\Delta_{2lpt}}{\Delta_{N-Body}} \right)^{1/3} \tag{3.5}$$

where Δ is the virial overdensity of a halo. For N-Body halos, Bryan & Norman (1998) fit

$$\Delta_{N-Body} = \frac{18\pi^2 + 82(\Omega_m(z) - 1) - 39(\Omega_m(z) - 1)^2}{\Omega_m(z)}, \tag{3.6}$$

where, for a flat ΛCDM cosmology,

$$\Omega_m(z) = \frac{\Omega_{m,0}(1+z)^3}{\Omega_{m,0}(1+z)^3 + (1 - \Omega_{m,0})}. \tag{3.7}$$

For the BOSS-LOWZ fiducial cosmology and simulation redshift this gives $\Delta_{N-Body} = 264$. To evaluate the same quantity for the 2LPT fields the relationship between the linear and non-linear overdensity, δ_{2lpt} and δ_L at second order is used

$$\Delta_{2lpt} = \delta_{2lpt} + 1 = \left(1 - \frac{\delta_L D_1}{3} - \frac{\delta_L^2 D_1^2}{21} \right)^{-3}. \tag{3.8}$$

D_1 is the linear growth factor. Using $\delta_L D_1$ = 1.68, which is the value recovered from the sphereical collapse approximation, $\Delta_{2lpt} = 35.4$. In turn this gives $b_{2lpt} \approx 0.39$ which is used as input for CM_HALOFINDER. Taking 10 particles as the minimum number required to constitute a halo, 500 sets of halos with $M_{halo} \geq 5 \times 10^{12} M_\odot h^{-1}$ were recovered and used for the subsequent creation of 1000 mock catalogues matching the clustering and survey geometry of the BOSS -LOWZ dataset.

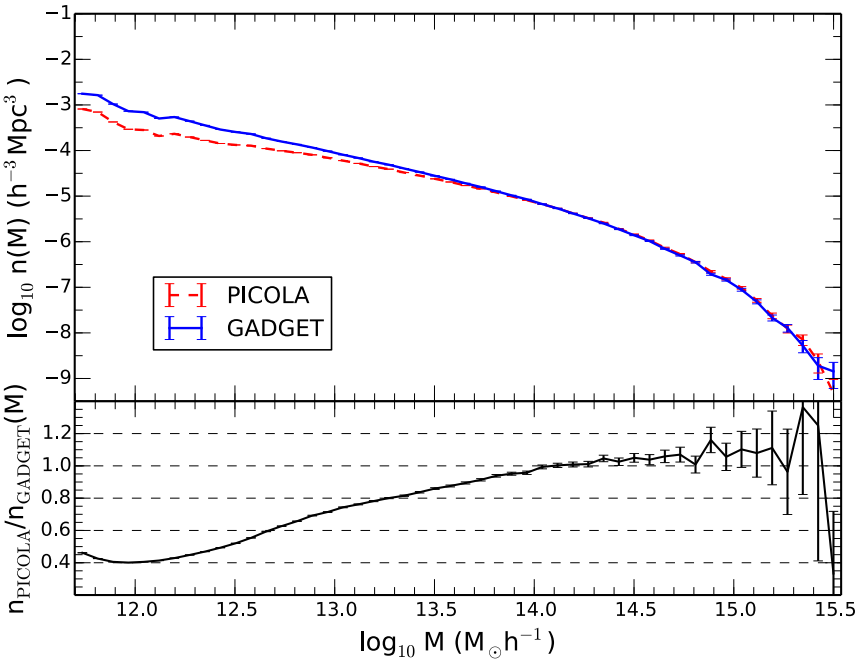


Figure 3.5: A comparison of the halo mass function from the MGS GADGET-2 and PICOLA simulations run from the same initial conditions. There is a slight lack of halos on small scales due to the finite mesh resolution, but this is easily compensated for with the HOD fitting described later.

3.3.3 Application to MGS Simulations

In comparison to the method used for the BOSS-LOWZ mocks, halos for the MGS L-PICOLA simulations were generated using a linking length equal to the commonly used value of $b = 0.2$. Because L-PICOLA can simulate small scale clustering much more accurately than 2LPT alone, a correction to the linking length is no longer required to recover the halo mass function with reasonable accuracy. This will be quantified later. The position and velocity of the centre-of-mass of each halo is calculated by averaging over all of the constituent particles of that halo. The halo mass, M, is given by the individual particle mass multiplied by the number of constituent particles that make up the halo. The virial radius is then estimated as

$$R_{vir} = \left(\frac{3M}{4\pi\rho_c(z)\Delta_{vir}\Omega_m(z)} \right)^{1/3}, \tag{3.9}$$

where $\rho_c \approx 2.77 \times 10^{11}\, h^2\, \mathrm{M}_\odot\, \mathrm{Mpc}^{-3}$ is the critical density, and we use a value $\Delta_{vir} = 200$ (e.g. Tinker et al. 2008).

Fig. 3.5 shows the level of agreement between halo mass functions recovered from L-PICOLA and GADGET-2 runs with the same simulation parameters and initial conditions. The difference in halo number density for low-mass halos is a direct consequence

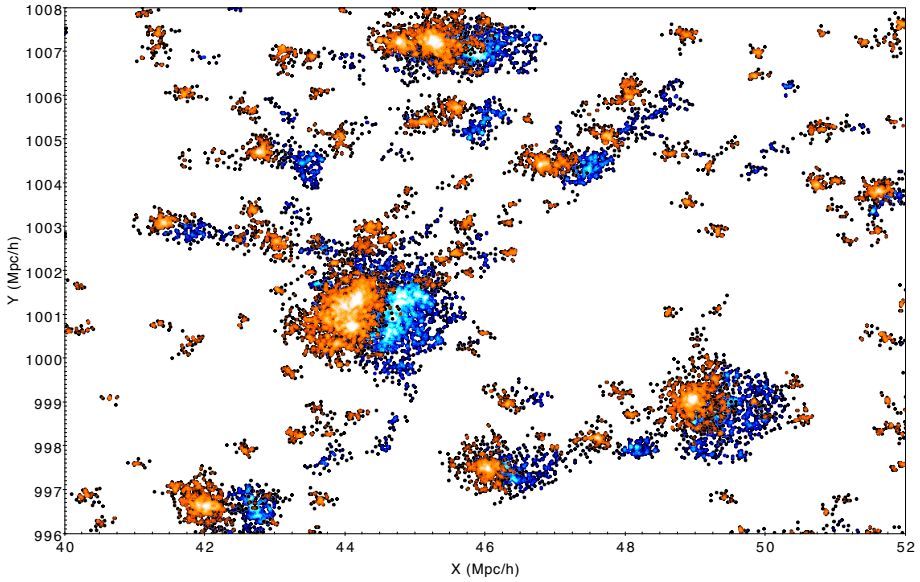


Figure 3.6: A visualisation of a small subset of the halos from the GADGET-2 (orange) and L-PICOLA (blue) simulations, where the constituent particles of the halos have been plotted. There is generally good agreement between the two fields, although the larger L-PICOLA halos are generally less compact than the corresponding GADGET-2 halos, and there is a lack of small halos compared to those from the fully non-linear simulation.

of the mesh resolution of the L-PICOLA dark matter simulations. As L-PICOLA does not calculate additional contributions to the inter-particle forces (i.e., via a Tree-level Particle-Particle summation) on scales smaller than the mesh, using instead the approximate, interpolated forces from the nearest mesh points, it does not quite produce the correct structure on the order of a few mesh cells or smaller. This results in slightly 'puffy' halos.

This is shown visually in Fig 3.6 where a small subset of the particles within halos are plotted for the GADGET-2 and L-PICOLA simulations. Both the larger extent of the L-PICOLA halos compared to their GADGET-2 counterparts and the lack of low mass L-PICOLA halos is evident, although the agreement between the two fields is generally quite remarkable.

As a more quantitative comparison, the normalised number of dark matter particles within halos for a given mass range is plotted in Figure 3.7 as a function of their separation from the centre of mass, normalised by the halo virial radius. For the halo mass range in question the constituent particles of the L-PICOLA halos are located at slightly larger radii than their GADGET counterparts. This difference is slightly reduced for higher mass halos where the overall properties of the halo are still captured, however some discrepancy does remain.

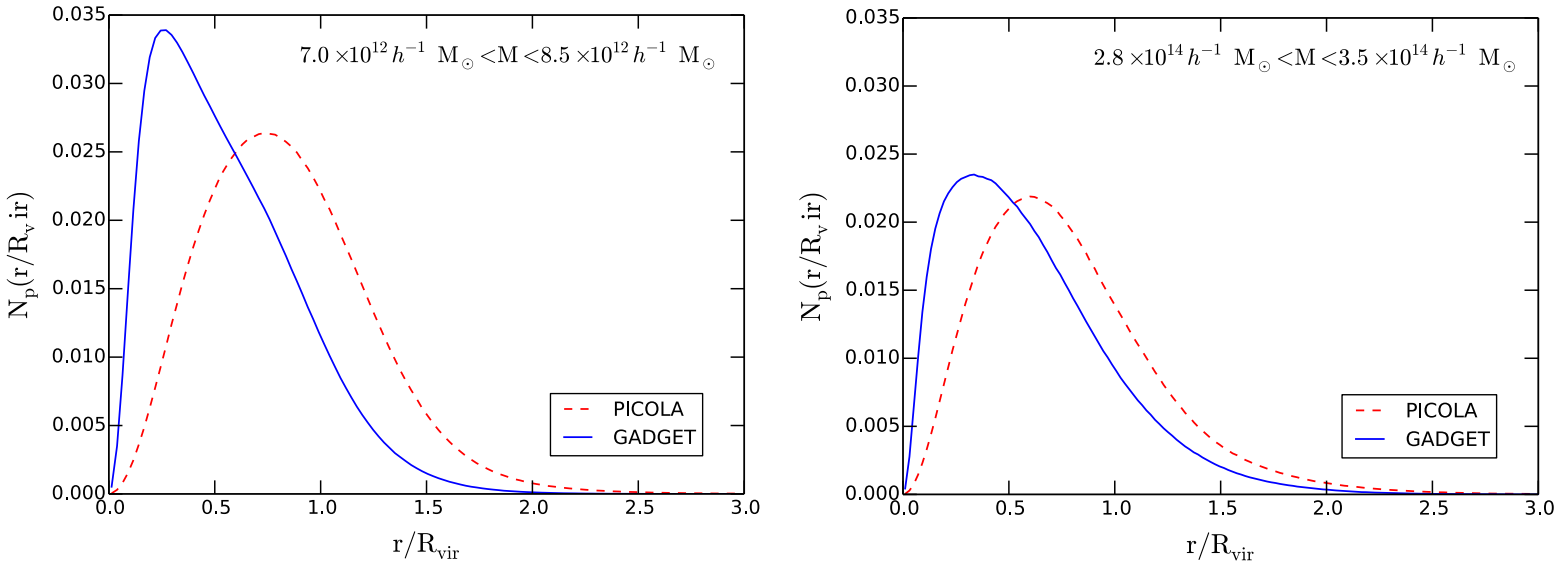


Figure 3.7: The normalised number of constituent dark matter particles found within an MGS halo as a function of their separation from the halo centre of mass, in units of the virial radius, for a given halo mass range. The halos from L-PICOLA are generally more dispersed than those from GADGET-2, where the particles have not collapsed sufficiently. This in turn leads to a slight lack of low mass halos overall, which is naturally corrected for in the HOD fitting method. The internal distribution of particles within halos of higher masses is generally closer to that expected for a fully non-linear simulation, but some discrepancy remains.

Regardless of this, the effect is small enough over the halo mass range of interest for the MGS that no correction is necessary before applying the HOD model. In addition, as described in Subsection 3.4, the HOD parameters are determined by directly populating the mock dark matter halos. The deficit of lower mass halos is thus compensated for by assigning galaxies to lower mass halos.

Comparison of Matched Halos

As a more complete comparison between the halos recovered from the L-PICOLA and GADGET-2 simulations, halos within the two runs were matched based on their member particles. For each GADGET-2 halo the constituent particles were identified and for each particle the corresponding L-PICOLA halo was found. The L-PICOLA halo that shared the most particles with the GADGET-2 halo is identified as the matching halo. However there are a few caveats. For many of the low mass GADGET-2 halos none of the constituent particles are within L-PICOLA halos (at least in the sense that halos are identified as bound structures containing more than 20 particles), hence these halos are not matched. Also, for a GADGET-2 halo whose constituent particles are found in equal numbers within *two* L-PICOLA halos, the halo closest to the centre-of-mass of the GADGET-2 halo was taken. Finally, this process can result in duplicate matches when multiple GADGET-2 halos are matched to the same larger L-PICOLA halo. This is due to the weaker non-linear collapse in the L-PICOLA dark matter field. Where there are distinct small mass halos around a much larger halo in the GADGET-2 run, the L-PICOLA simulation contains a larger mass halo has not fully collapsed and hence the FoF algorithm connects both the massive halo and surrounding smaller halos into a single structure. This is the same effect that creates 'puffy' halos as mentioned previously. For the purposes of halo matching we simply take the largest GADGET-2 halo for a set of L-PICOLA duplicates. We would expect this then to mean that for some of the largest matched halos, the L-PICOLA halo mass is slightly larger than its GADGET-2 counterpart. Overall, of the $\sim 20,000,000$ GADGET-2 halos, $\sim 6,700,000$ are identified as having unique L-PICOLA matches.

Fig. 3.8 shows the halo mass function for all the matched halos. One can see that indeed, at the high mass end there seems to be a slight excess in the L-PICOLA mass function compared to that measured from the GADGET-2 run, due to the inability of L-PICOLA to fully model the collapse of structure and consequently creating a single large halo where there should be a massive halo surrounded by identifiably separate small halos. There is also some discrepancy at the low and intermediate masses $M_{halo} < 10^{14}h^{-1}M_{\odot}$, where there is an increasingly large excess of GADGET-2 halos before a sharp decrease in the mass function compared to the L-PICOLA halos. This feature arises due to the smaller number of low mass L-PICOLA halos, and from the fact that for intermediate mass halos,

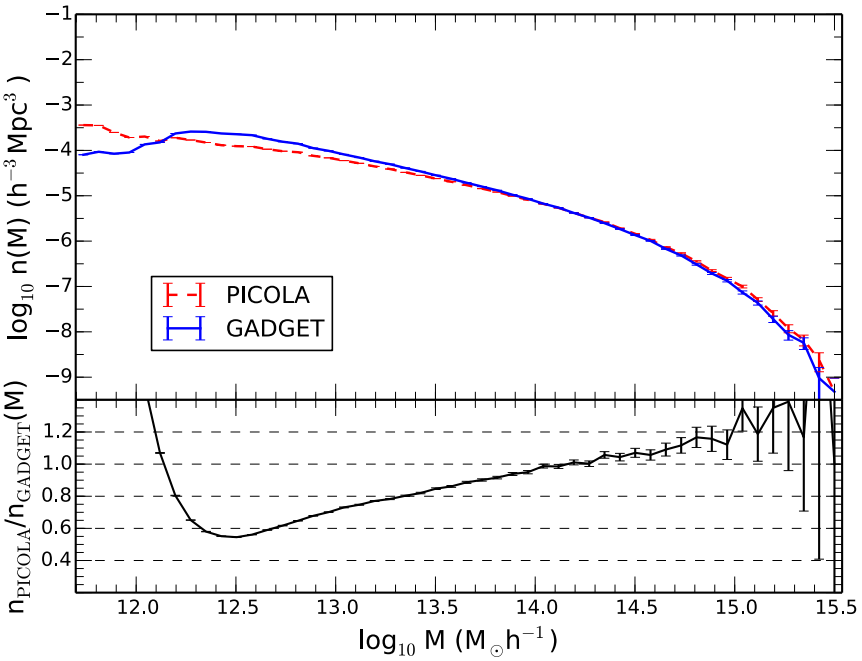


Figure 3.8: A comparison of the halo mass function for matched halos from the MGS GADGET-2 and PICOLA simulations run from the same initial conditions. The inaccuracies in non-linear collapse of the L-PICOLA dark matter result in higher masses for the largest matched halos where a single halo in the L-PICOLA simulation corresponds to separate small halos surrounding a much larger halo in the GADGET-2 run. Consequently the deficit of low mass L-PICOLA halos results in few matches to the low mass GADGET-2 halos, and hence these are missing from the matched subsample, whilst for intermediate masses the GADGET-2 halos are matched to lower mass L-PICOLA halos . Overall this results in a GADGET-2 mass function for matched halos that is higher at intermediate masses and lower at small masses than its L-PICOLA counterpart.

the halo identified in the L-PICOLA simulation is typically smaller than that identified in GADGET-2 as some of the dark matter in the L-PICOLA simulation has not collapsed enough to be included by the FoF algorithm. This latter effect causes an increase in the GADGET-2 mass function for matched halos, whilst the former means that many of the low mass GADGET-2 halos have no match and so are not included in the sample, causing the sharp drop at low masses.

These effects are also apparent in the clustering of the matched and unmatched halos with masses $10^{12}\,h^{-1}M_{\odot} < M_{halo} < 10^{13}\,h^{-1}M_{\odot}$. Fig. 3.9 shows the power spectrum measured from all the GADGET-2 halos within this mass range, those that are matched to the L-PICOLA simulation and those that have no matches. Although the clustering of all three samples has a similar shape, the clustering of the matched low mass halos is of lower amplitude than that of the full sample, whilst the clustering of unmatched halos is higher. This is because the unmatched low mass halos are generally located in the vicinity of a much larger halo, at higher pekas in the density field, and hence are more strongly clustered than is typical of halos with that mass. Conversely, the matched halos are more often found in less dense regions and so are more weakly clustered. This reflects that fact that many of the low mass GADGET-2 halos located near larger halos do not have a corresponding L-PICOLA halo as the large halo in the approximate simulation is less compact than it should be, and as such the FoF algorithm includes the smaller mass halos within this single structure.

3.4 Assigning Galaxies to Halos

The MGS is approximately volume limited to $z < 0.17$ (due to the $M_r < -21.2$ restriction), above which the number density drops due to the $r_{\rm pet} < 17.6$ magnitude limit. The $n(z)$ is large enough that the sample is cosmic-variance limited ($n(z)P(k) > 1$) over the entire redshift range for $k < 0.26\,h\,{\rm Mpc}^{-1}$. Furthermore, the $n(z)$ is constant to within a factor of two, making it more constant than the BOSS 'CMASS' sample that has been modelled as having a single HOD at its effective redshift in cosmological analyses (e.g., Anderson et al. 2014a,b; Manera et al. 2013).

The MGS halos are populated in a very similar way to that of Manera et al. (2013) using the HOD model (Berlind & Weinberg, 2002). Within this framework galaxies are assigned to halos based solely on the mass of the halo, with a conditional probability, such that a halo of mass M has a probability $P(N|M)$ of containing N galaxies. Models for how the galaxies are distributed within the halo and their velocities must also be assumed. It is common to refine this further by splitting the galaxies into central and satellite types, the distributions of which are treated independently. Two mass-dependent functions are defined, $\langle N_{cen}(M)\rangle$ and $\langle N_{sat}(M)\rangle$, where $\langle N_{cen}(M)\rangle$ denotes the probability that a

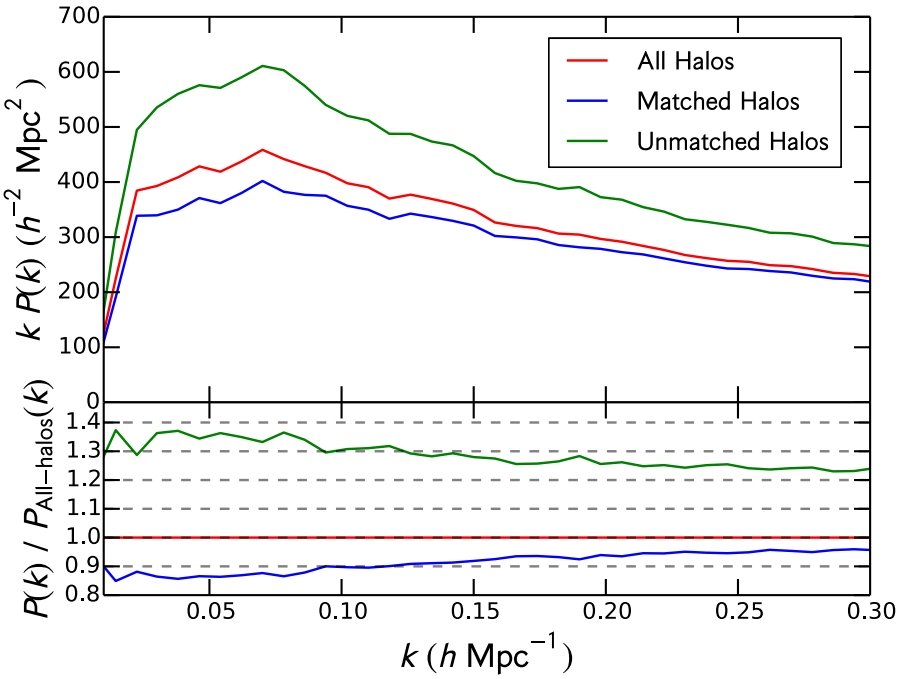


Figure 3.9: A comparison of the power spectra measured from the GADGET-2 simulation for all halos with $10^{12}\,h^{-1}M_{\odot} < M_{halo} < 10^{13}\,h^{-1}M_{\odot}$, further divided into those halos that have been matched to counterparts in the L-PICOLA simulation, and those that have no identified match. The bottom panel shows the ratio of the power spectra to that measured from the full simulation. The unmatched halos have a greater clustering amplitude than the full sample, whilst the matched halos have a lower amplitude. This is because the unmatched halos are more often found in the vicinity of larger halos, at high peaks in the density field, and hence have stronger clustering. The opposite is true for the matched halos.

halo of mass M contains a central galaxy and $\langle N_{sat}(M)\rangle$ is the mean of the poisson distribution from which the number of satellite galaxies is randomly generated. These functions are determined from a fit to the MGS data, by iterating over the following steps:

1. Populate a subset of the mocks using a given set of HOD parameters.
2. Mask the mock galaxies so that they match the data.
3. Subsample the mock galaxies to match the idealised $n(z)$.
4. Calculate the average power spectrum of our populated mocks and compare to the data.

10 mocks are used to fit the HOD. These are populated and masked individually, but the radial selection function is reproduced by subsampling based on the ratio between the analytic fit to the data $n(z)$ and the average $n(z)$ of the 10 mocks. The fit is performed using a downhill simplex minimisation of the χ^2 difference between the average, 10-mock power spectrum and number density, and the data power spectrum and fitted number density formula. The fit is performed twice. First, analytic errors on the power spectrum from Tegmark (1997) are used, where for a power spectrum, $P(k_i)$, averaged over a bin of width Δ_{k_i} centred at k_i,

$$C_{i,j} = \frac{4\pi^2 P^2(k_i)}{k_i^2 \Delta_{k_i} V_{eff}(k_i)} \delta^D(k_i - k_j), \tag{3.10}$$

where $\delta^D(k_i - k_j)$ is the Dirac delta function and

$$V_{eff}(k) = \int d^3\boldsymbol{r} \left(\frac{n(\boldsymbol{r})P(k)}{1 + n(\boldsymbol{r})P(k)} \right)^2. \tag{3.11}$$

This analytic form of the covariance matrix holds under the assumption that the underlying density field can be approximated as a Gaussian random field and where terms proportional to n^{-2} and n^{-3} are subdominant and can be neglected. See Chapter 5 for a detailed derivation of this.

The second fit to the data is performed using the covariance matrix from the first fit and is used as the final best fit model. Each of the individual HOD fitting steps are detailed below.

3.4.1 Populating the Mocks Using the HOD

The five parameter functional form of Zheng et al. (2007) was used for the $\langle N_{cen}(M)\rangle$ and $\langle N_{sat}(M)\rangle$ functions, where,

$$\langle N_{cen}(M)\rangle = \frac{1}{2}\left[1 + \text{erf}\left(\frac{\log M - \log M_{min}}{\sigma_{\log M}}\right)\right],$$

$$\langle N_{sat}(M)\rangle = \langle N_{cen}\rangle \left(\frac{M - M_{cut}}{M_1}\right)^{\alpha}. \tag{3.12}$$

Here $\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$ is the error function and $M_{min}, \sigma_{\log M}, M_{cut}, M_1$ and α are five free parameters, which must be fit. For a halo of $M < M_{cut}$, $\langle N_{sat}\rangle = 0$ is enforced manually and in the case where a satellite galaxy but no central galaxy is assigned to a halo, one of the potential satellite galaxies is removed and replaced with a central.

Galaxy Positions

Central galaxies are placed at the centre of mass of the halo, and satellites at radii $r \leq R_{vir}$ with probability derived from the NFW profile (Navarro et al., 1996)

$$\rho(r) = \frac{4\rho_s}{\frac{r}{r_s}\left(1 + \frac{r}{r_s}\right)^2}, \tag{3.13}$$

where $r_s = R_{vir}/c_{vir}$ is the characteristic radius, at which the slope of the density profile is -2, and ρ_s is the density at this radius. c_{vir} is the concentration parameter, which is calculated for a halo of mass M using the fitting formulae of Prada et al. (2012). These fitting formulae are derived from fits to multiple N-Body simulations. For a halo of mass M at redshift z

$$c_{vir}(M,z) = A\frac{c_{min}(x)}{c_{min}(1.393)}\left[\left(\frac{\sigma_{min}(x)\sigma(M,x)}{b\sigma_{min}(1.393)}\right)^c + 1\right]\exp\left(\frac{d\sigma^2_{min}(1.393)}{\sigma^2_{min}(x)\sigma^2(M,x)}\right), \tag{3.14}$$

where $A = 2.881$, $b = 1.257$, $c = 1.022$ and $d = 0.060$ are constants and

$$x = \left(\frac{\Omega_{\Lambda,0}}{\Omega_{m,0}}\right)^{1/3}\frac{1}{(1+z)^3} \tag{3.15}$$

partially incorporates the cosmological dependence of this fitting function. The remaining cosmological dependence is incorporated into $\sigma(M,x)$, the rms density fluctuations within the simulations. Klypin et al. (2011) fit this as

$$\sigma(M,x) = D_1(x)\frac{16.9y^{0.41}}{1 + 1.102y^{0.20} + 6.22y^{0.333}}, \tag{3.16}$$

where $y = (10^{12}\mathrm{M}_{\odot}h^{-1}/M)$, is the inverse normalised halo mass and $D_1(x)$ is the linear growth factor, rewritten in such a way as to absorb the redshift dependence into the parameter x,

$$D_1(x) = \frac{5}{2}\left(\frac{\Omega_{m,0}}{\Omega_{\Lambda,0}}\right)^{1/3}\frac{\sqrt{1+x}}{x^{3/2}}\int_0^x \frac{t^{3/2}}{(1+t^3)^{3/2}}dt \tag{3.17}$$

Finally the two remaining functions c_{min} and σ_{min} are the minima of the concentration and rms density fluctuations. These are calculated via

$$c_{min}(x) = c_0 + (c_1 - c_0)\left[\frac{1}{\pi}\arctan(\alpha(x - x_0)) + \frac{1}{2}\right] \tag{3.18}$$

$$\sigma_{min}(x) = \sigma_0 + (\sigma_1 - \sigma_0)\left[\frac{1}{\pi}\arctan(\beta(x - x_1)) + \frac{1}{2}\right] \tag{3.19}$$

where $c_0 = 3.681$, $c_1 = 5.033$, $\alpha = 6.948$, $x_0 = 0.424$, $\sigma_0 = 1.047$, $\sigma_1 = 1.646$, $\beta = 7.386$, and $x_1 = 0.526$ are all constants.

On top of this we add a dispersion to the mass-concentration relation using a lognormal distribution with mean equal to that evaluated from the fitting functions and variance $\sigma = 0.078$. This is the same value as that used in Manera et al. (2013) and is a typical value, as measured from fitting NFW profiles to halos recovered from simulations (Giocoli et al., 2010).

Galaxy Velocities

Both central and satellite galaxies are given the velocity of the centre of mass of the halo. Satellite galaxies are then assigned an extra peculiar velocity contribution drawn from a Gaussian, with the velocity dispersion calculated from the virial theorem

$$\langle v^2 \rangle = \left\langle \frac{GM(r)}{r} \right\rangle. \tag{3.20}$$

For an NFW profile, the mass inside a radius r is

$$M(r) = 4\pi\rho_s r_s^3 \left[\ln\left(\frac{r_s + r}{r_s}\right) - \frac{r}{r_s + r}\right], \tag{3.21}$$

and hence the velocity dispersion for a halo of mass M is

$$\langle v^2 \rangle = \frac{GM}{r_s}\frac{c(1+c) - (1+c)\ln(1+c)}{2((1+c)\ln(1+c) - c)^2}. \tag{3.22}$$

In order to assign the additional satellite velocities in each direction a gaussian distribution with zero mean and variance $\langle v^2 \rangle/3$ was used.

Redshift-Space Distortions

To simulate the effects of Redshift-Space Distortions each galaxy is displaced along the line-of-sight by

$$\Delta s_{los} = \frac{v_{los}}{H(z)a}, \tag{3.23}$$

Given Δs_{los} and a galaxy's true position, angles and redshifts are determined using the fiducial cosmology, placing the observer at the centre of each simulation box.

3.4.2 Masking the Mocks

Masking of the mocks begins by converting each galaxy's cartesian (x, y, z) coordinates to RA, DEC and redshift using the transformation,

$$\text{RA} = \frac{180}{\pi}\arctan\left(\frac{y}{x}\right) + 180 \tag{3.24}$$

$$\mathrm{DEC} = \frac{180}{\pi}\arcsin\left(\frac{z}{r}\right) \tag{3.25}$$

where $r^2 = x^2 + y^2 + x^2$. The well-known ATAN2 routine is used to compute the arctan function and ensure consistency in treatment of the sign of each galaxy's $x-$ and y-coordinates. The redshift of each galaxy is computed using a spline-fit to a lookup table of redshift against comoving distance created using the fiducial cosmology.

Full mock catalogue sample

During creation of the full mock catalogue sample, the public code MANGLE (Swanson et al., 2008) is used to mask the galaxies and apply subsampling based on the completion in each region of the survey. The angular footprint of the MGS, and by design the mocks, is displayed in blue in Fig. 3.10. The red patch in Fig. 3.10 shows the angular footprint of the MGS after rotating the coordinates via

$$RA \Rightarrow RA + \pi, \tag{3.26}$$

$$DEC \Rightarrow -DEC, \tag{3.27}$$

and once again applying the mask. The dark matter simulations are large enough to cover the full-sky out to $z = 0.2$ and so can be used to create multiple mock galaxy catalogues that match the MGS footprint, reducing the noise in the estimate of the covariance matrix at almost no extra cost. In principle, one could fit ~ 6 replicates of the survey in each full-sky simulation without overlap, though not, perhaps, without significant cross-correlation between patches taken from the same realisation. In practice, for simplicity, two survey patches were generated from each simulation.

During HOD fitting

Unlike for the full ensemble of mock catalogues, during fitting of the HOD, for the sake of expediency, only a single realization is cut from the mock galaxy field and instead of using MANGLE an approximate mask is used. To create this approximate mask, a list of equally spaced cartesian coordinates is generated on a grid and the corresponding RA, DEC and redshifts computed. These are the passed to MANGLE and a weight is assigned to each grid point based on survey completion. For regions outside the survey mask the completion is 0. For ease the grid upon which the cartesian coordinates are generated is chosen to be the same extent and resolution as that used to ultimately compute the power spectrum of the mocks and data during HOD fitting. The mock galaxies are then masked by comparing to this grid. Overall this creates a mask accurate to some smoothing scale, which is expected to have negligible effect on the best-fitting HOD model returned by the fitting procedure.

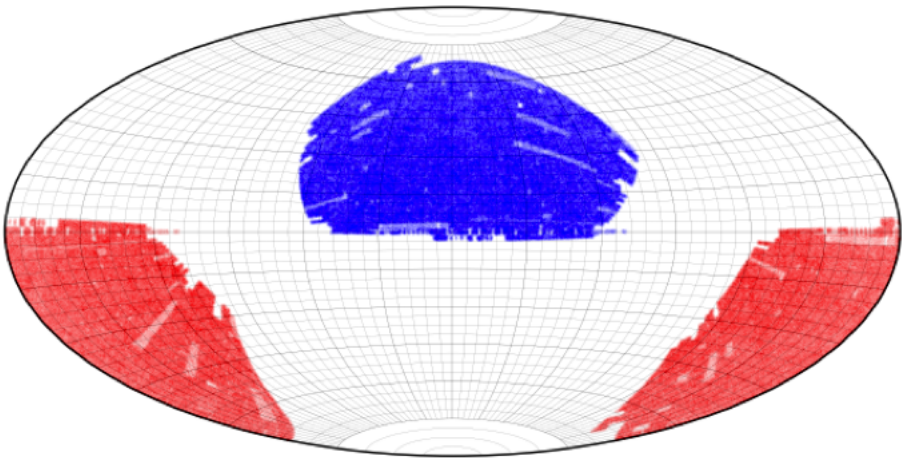


Figure 3.10: The blue area shows a flat, all-sky projection of the footprint of the MGS sample, which occupies 6,813 deg^2. The red area shows the same geometry, after a 180^o rotation. This illustrates how two mock galaxy samples are produced from every full-sky dark matter halo catalogue.

3.4.3 Subsampling the Mocks

The third stage in the HOD fitting procedure is subsampling the mocks based on the analytic $n(z)$ model. Although the MGS sample is approximately volume limited up to $z = 0.17$, the $n(z)$ is not quite flat up to this redshift due to the lack of dust-extinction correction in the measured galaxies themselves. The number density also drops off significantly after $z = 0.17$ and as a single HOD will return a flat number density it is necessary to subsample the mocks to match the expected number density as a function of redshift, both when fitting the HOD and when producing the final ensemble of mock catalogues.

Each set of mock galaxies should be treated as a separate realization of some underlying distribution, and as such each realization may have an $n(z)$ that differs slightly from the analytic model due to clustering along the line of sight and shot noise. Subsampling each mock individually can result in removing signal along the line of sight and underestimating the shot noise component of the number density. As such, during HOD fitting and application of the final best fit model, the MGS mocks are subsampled as an ensemble, i.e., the average number density is subsampled to match the analytic fit and each mock is subsampled by the same amount. This means that any mocks with a natural excess of galaxy clustering in the radial direction will retain this.

3.4.4 Calculating the Power Spectrum

The monopole moment of the power spectrum is used to determine the best-fitting HOD model for the mocks, as it is faster to compute than its configuration-space analogue

and the corresponding covariance matrix can be easily approximated. The monopole of the power spectrum, denoted $P(k)$, is computed using the method of Feldman et al. (1994). The overdensity is computed on a grid containing 1024^3 cells in a box of edge length $2000\,h^{-1}\,\mathrm{Mpc}$. This provides ample room to zero pad the galaxies to improve the frequency sampling, and results in a Nyquist frequency of $1.6\,h\,\mathrm{Mpc}^{-1}$, much larger than the largest frequency of interest. Galaxies and randoms are weighted based on the number density as a function of redshift,

$$w_{FKP}(z) = \frac{1}{1 + n(z)P_{FKP}} \tag{3.28}$$

where $P_{FKP} = 16000\,h^{-3}\mathrm{Mpc}^3$ was used for the MGS sample. After Fourier transforming the overdensity grid the spherically-averaged power spectrum is calculated in bins of $\Delta k = 0.008\,h\,\mathrm{Mpc}^{-1}$, correcting for gridding effects and shot-noise. The HOD is fit using scales $0.02\,h\,\mathrm{Mpc}^{-1} \leq k \leq 0.3\,h\,\mathrm{Mpc}^{-1}$, which are the scales over which one would typically fit the BAO and RSD signals. Lower than this and the covariance matrix becomes dominated by the window function and noise due to the finite number of modes that can be sampled. At larger k non-linear effects dominate, which are difficult to model when fitting the BAO and RSD. The power spectrum of the MGS data is displayed as points in Fig. 3.11. The smooth curve and error-bars display the mean of the mock $P(k)$ and their standard deviation.

3.4.5 Best-fitting HOD

After following the fitting procedure described previously, the best-fit HOD for the MGS is found to be

$$\begin{aligned}
M_{min} &= 1.51 \times 10^{13}\,h^{-1}\mathrm{M}_\odot, \\
M_{cut} &= 1.41 \times 10^{13}\,h^{-1}\mathrm{M}_\odot, \\
M_1 &= 8.71 \times 10^{13}\,h^{-1}\mathrm{M}_\odot, \\
\sigma_{logM} &= 0.904, \\
\alpha &= 1.18, \\
n &= 7 \times 10^{-4}\,h^3\,\mathrm{Mpc}^{-3},
\end{aligned}$$

where n is dependent on the five other parameters. These HOD parameters are in good agreement with the HOD parameters reported by Zehavi et al. (2011) for another SDSS galaxy sample with similar number density and magnitude limit. Fig. 3.12 shows the percentage difference between the average mock power spectrum and the power spectrum of the data. The errors come from the covariance estimated from the full, masked, mock sample. The amplitude of the power spectra matches well on all scales, with $\sim 5\%$

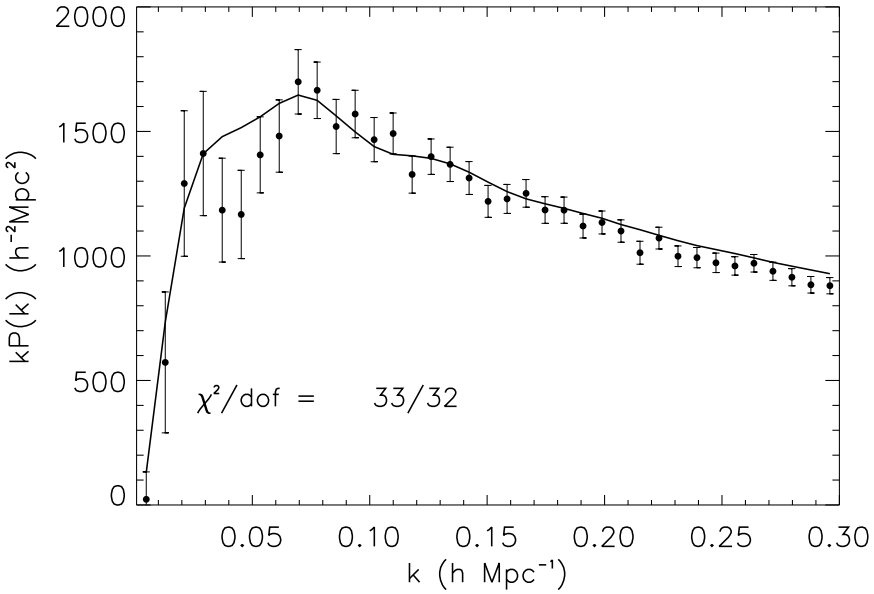


Figure 3.11: The power spectrum of the MGS. Points show the data and the solid line shows the mean of the mocks. The error bars come from the diagonal elements of the covariance matrix constructed using the mock catalogues.

agreement up to $k = 0.3\,h\,\mathrm{Mpc}^{-1}$. On the largest scales the window function has a large effect, correlating the data and making offsets look by eye to be more significant they they are. This is reflected in the χ^2 of the fit, as $\chi^2 = 33$ for 32 degrees of freedom (37 k-bins and 5 free parameters) indicating that the fit is good.

Fig. 3.13 shows the expected number of galaxies in the mock halos for the best fit HOD model. This highlights how the clustering properties of the data are recovered so accurately, even though the L-PICOLA simulations lack the correct number of low mass halos. All of the satellite galaxies exist in halos with $M > 10^{13}h^{-1}\mathrm{M}_{\odot}$, which are well recovered by the simulations. Below this mass, where the simulations lack sufficient number density, the probability of finding any galaxies within a halo also drops rapidly, such that even though these halos are more abundant in general, the contribution to the total clustering from these halos is small in comparison to the larger mass halos.

There exists significant degeneracy between the five free HOD parameters, which cannot be broken completely by just the one-dimensional, two-point clustering statistics. Three-point statistics could be used to break this degeneracy (Kulkarni et al., 2007), however this would be prohibitively time-consuming and potentially very noise dominated. Another possibility is to use the quadrupole or hexadecapole moments of the power spectrum, as these contain additional information about the position and velocity distribution of the satellite galaxies within their host halos (Hikage, 2014). Again, however, in the

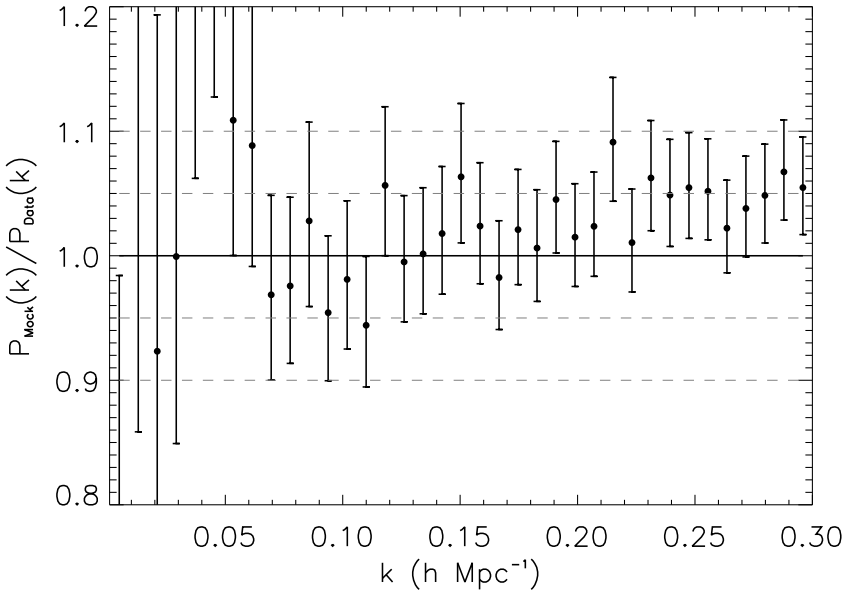


Figure 3.12: The percentage difference between the average mock power spectrum and that of the MGS data, with errors derived from the covariance matrix of the 1000 mock catalogues. There is good agreement ($\sim 5\%$) between these up to $k = 0.3\,h\,\mathrm{Mpc}^{-1}$ except on large scales (small k) where the window function introduces additional covariance between different k-bins.

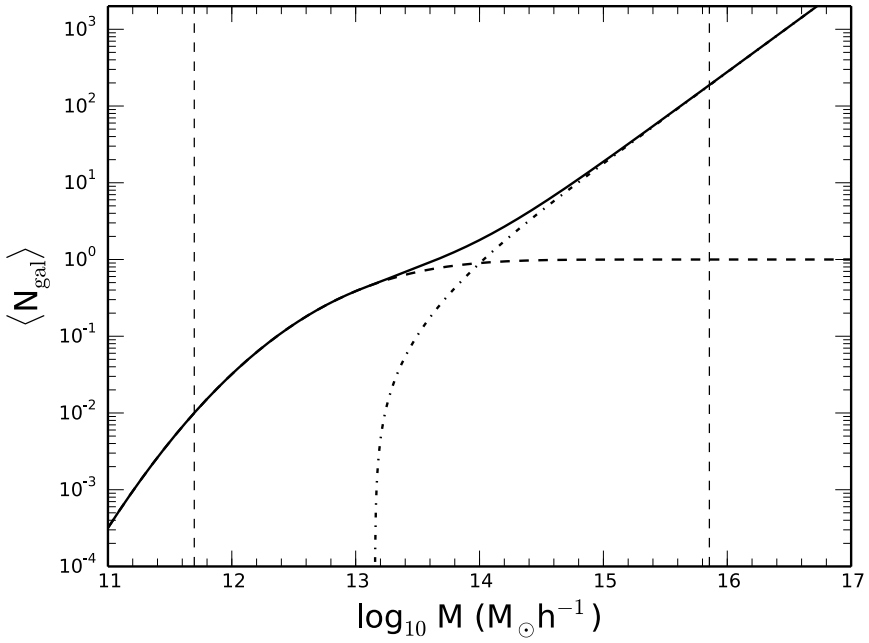


Figure 3.13: The expected number of galaxies in a halo as a function of halo mass for the best-fitting HOD parameters. The dashed line shows the probability of the halo hosting a central galaxy, and the dot-dashed line shows the average number of satellite galaxies within such a halo. The two vertical dashed lines denote the maximum and minimum halo masses across all 1000 mock catalogues.

case of the MGS these statistics will almost certainly be noise dominated. As such we leave these as future improvements for the mock catalogue production process.

3.5 Clustering of the MGS Mock Catalogues

3.5.1 Correlation Function

Even though the power spectrum and correlation function form a Fourier pair, because the mock HOD is fit to the power spectrum it is not guaranteed that the mock catalogues will have the same level of agreement with the MGS data in configuration space. This agreement is investigated using the minimum variance estimator of Landy & Szalay (1993), with galaxy and random weights as given in Eq. (3.28). The 2-D correlation function is calculated for the 1000 mock ensemble and the data by binning in the redshift space separation, s, and the cosine of the angle, μ, between the vector joining each pair of galaxies and the line of sight bisecting this. This uses bins of width $\Delta s = 1.0\,h^{-1}\,\mathrm{Mpc}$ and $\Delta\mu = 0.01$ for $0 < s \leq 200$ and $0 \leq \mu \leq 1$.

The multipole expansion of the two-dimensional correlation function is computed via the Riemann sum

$$\frac{2\xi_\ell(s)}{2\ell+1} = \sum_{i=1}^{100} 0.01\xi(s,\mu_i)P_\ell(\mu_i), \tag{3.29}$$

where $\mu_i = 0.01i - 0.005$ and $P_\ell(\mu)$ are the Legendre Polynomials of order ℓ. The monopole and quadrupole are then generated for different bin widths by re-summing the pair counts before applying Eq. (3.29).

Figs. 3.14 and 3.15 show the monopole and quadrupole of the correlation function for the average of the MGS mocks and for the data for the 24 measurements in the range $8 < s < 200h^{-1}$Mpc. The mean of the mock ξ_0 and ξ_2 do not match the data within the error-bars at many scales. However, only the diagonal elements of the covariance matrix are plotted, and the off-diagonal elements represent a significant component (see Fig. 3.17). A better comparison is the χ^2 between the mean of the mocks and the data, using the full covariance matrix. For both ξ_0 and ξ_2 the χ^2/d.o.f is slightly less than one, implying the anisotropic clustering in the mock samples is a good representation of the data, even down to $10h^{-1}$Mpc scales (and hence 'χ by eye' is a bad idea).

3.5.2 Covariance Matrix

We use the sample of N mock galaxy catalogues to estimate the covariance matrix for both the power spectrum and correlation function via

$$C_{i,j} = \frac{1}{N-1}\sum_N (x_i^N - \mu_i)(x_j^N - \mu_j) \tag{3.30}$$

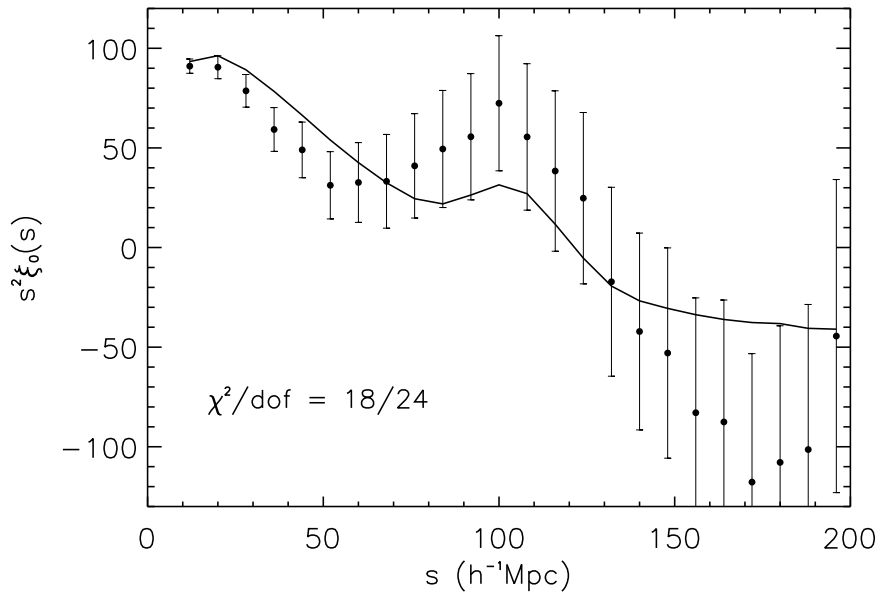


Figure 3.14: The monopole moment of the correlation function of the MGS. The solid line shows the mean of the mocks and the error bars come from the diagonal elements of the covariance matrix calculated from the 1000 mock realisations.

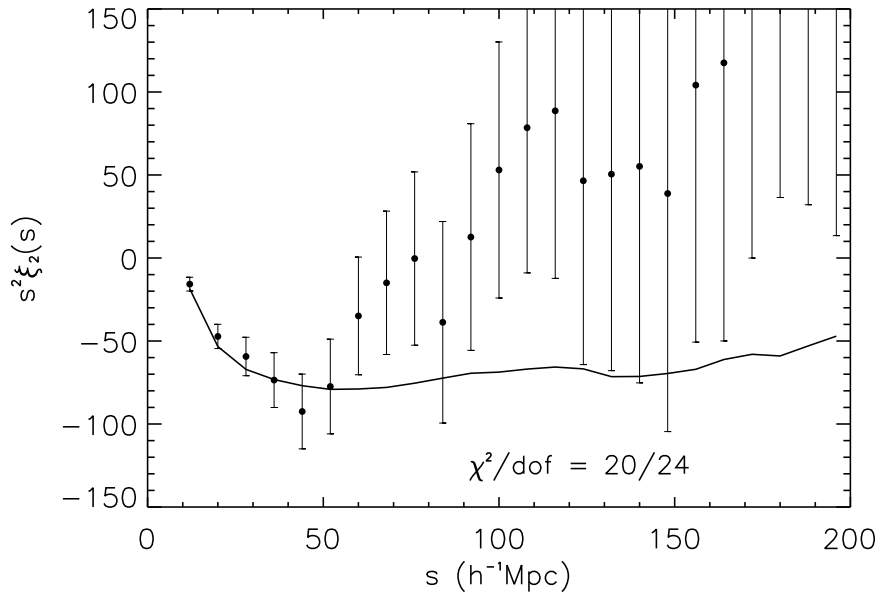


Figure 3.15: The quadrupole moment of the correlation function of the MGS and the mean of the mock galaxy catalogues. Though the agreement by eye looks poor on large scales, there exists significant covariance between the points at different scales, such that the chi-squared between the data and mocks is small.

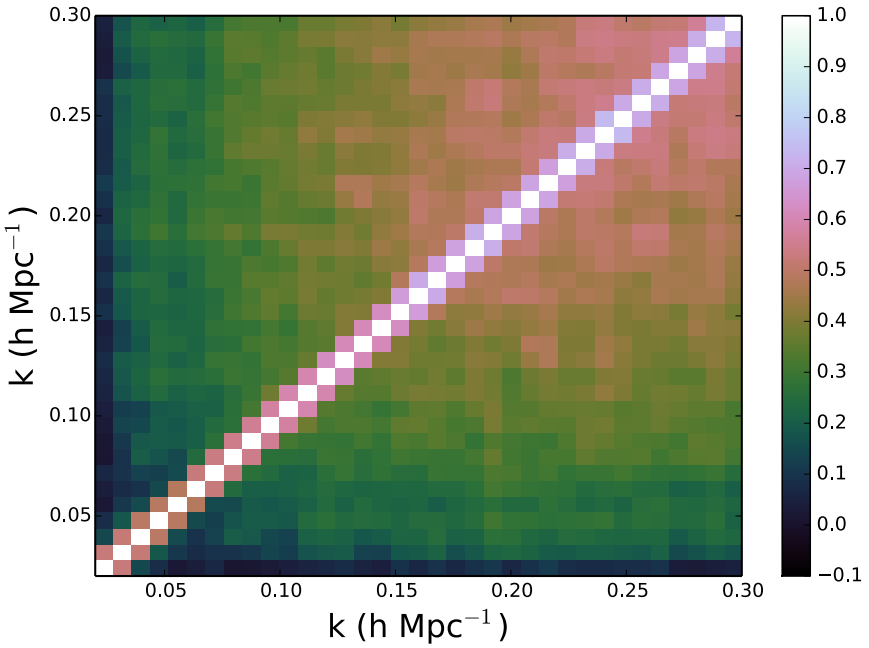


Figure 3.16: The power spectrum correlation matrix generated from the 1000 MGS mock catalogues between $k = 0.02\,h\,\mathrm{Mpc}^{-1}$ and $k = 0.3\,h\,\mathrm{Mpc}^{-1}$ and in bins of $\Delta k = 0.008\,h\,\mathrm{Mpc}^{-1}$.

where

$$\mu_i = \frac{1}{N}\sum_N x_i^N \tag{3.31}$$

is the average of the statistic x, measured from each mock.

Figs. 3.16 and 3.17 show the correlation matrix, $C_{i,j}^{red} = C_{i,j}/\sqrt{C_{i,i}C_{j,j}}$, for the power spectrum and the monopole and quadrupole moments of the correlation function using the fiducial binning scheme. There is significant off-diagonal covariance in the correlation function and non-negligible cross-covariance between the monopole and quadrupole, however the power spectrum covariance matrix is much more diagonal.

Estimates of the covariance matrix using a finite number of realisations are subject to errors based on the number of realisations used and the number of bins for which the covariance is being estimated. When fitting models to data one requires the 'true' covariance matrix to recover the correct likelihood function. The estimated covariance however is drawn from a Wishart distribution based on the 'true' covariance. This distribution is such that the whilst the covariance matrix estimated using Eq.3.30 is unbiased, its inverse ψ, which is used for likelihood analysis, is not. Furthermore, when the estimated inverse covariance is used to estimate the likelihood in lieu of the true inverse covariance matrix, one should marginalise over the probability of drawing a given inverse covariance ma-

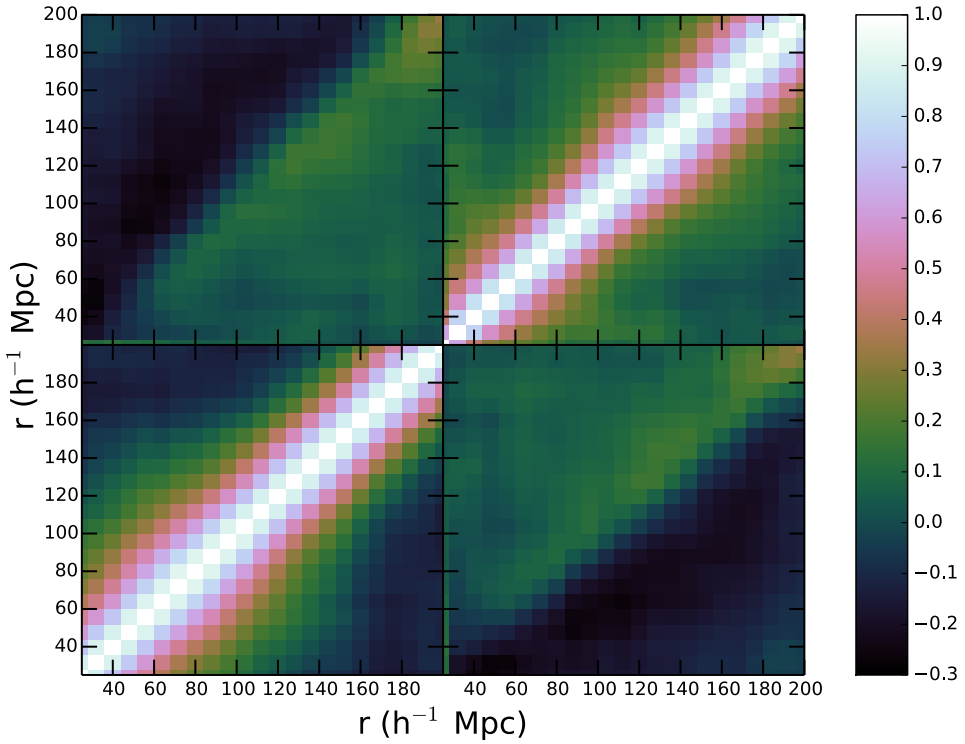


Figure 3.17: The correlation matrix for the correlation function monopole (bottom left) and quadrupole (top right) and the cross covariance between the two (top left/bottom right), in bins of $8\,h^{-1}\,\mathrm{Mpc}$ in the range $25\,h^{-1}\,\mathrm{Mpc} \leq s \leq 200\,h^{-1}\,\mathrm{Mpc}$.

trix based on the inverse Wishart distribution. This increases the width of the likelihood function and hence the errors on any measurements obtained.

Hartlap et al. (2007) provide a correction to the inverse covariance matrix to account for the first of these effects

$$\psi = \left(1 - \frac{n_b + 1}{n_s - 1}\right) C^{-1}, \tag{3.32}$$

where n_b is the number of bins, n_s is the number of mock realisations and C^{-1} is the estimated inverse covariance matrix. For 1000 mocks and the binning schemes used in this and the next chapter, this is a corrective factor of 1.038 and 1.025 for the power spectrum and correlation function respectively.

Building on the work of Dodelson & Schneider (2013), Percival et al. (2014) provide a correction for the second of these effects. This takes the form of a multiplicative constant based on the number of measurement bins, mocks and free parameters being fit n_p, which should be applied to any likelihoods or standard deviations (these get multiplied by the square root of the corrective factor) calculated using the estimated inverse covariance matrix,

$$m_\sigma = \frac{1 + B(n_b - n_p)}{1 + 2A + B(n_p + 1)}, \tag{3.33}$$

where

$$A = \frac{1}{(n_s - n_b - 1)(n_s - n_b - 4)}, \tag{3.34}$$

$$B = A(n_s - n_b - 2). \tag{3.35}$$

For example, a number of bins $n_b = 34$ and fitting parameters $n_p = 8$, gives only a small correction to the inverse covariance matrix of $m_\sigma = 1.017$. This is the value used in the next chapter when fitting the RSD signal in the MGS data.

If one is fitting to the same distribution from which the covariance is calculated, i.e. fitting to the mean of the mocks, the corrective factor must be modified slightly to

$$m_\nu = m_\sigma \frac{n_s - 1}{n_s - n_b - 2}. \tag{3.36}$$

Using the same parameters as for m_σ this gives $m_\nu = 1.054$. Where appropriate, these corrections have been applied to all analyses in this and the following chapter. However, the factors are small enough, due to the large number of mocks used, that their effect is minimal.

3.6 Systematic Tests

3.6.1 Independence of Mocks

The coordinate transformation that allows two distinct MGS mocks to be created from each dark matter realisation puts the two patches as far apart as possible to minimise the covariance between mocks based on the same dark matter cube. The minimum possible distance between two objects in different patches is $170\,h^{-1}\,\mathrm{Mpc}$. Whilst this is within the range of scales of interest for BAO and RSD measurements, the total cross-correlation between patches is very small. The mocks are numbered such that pairs of mocks (e.g. 1 & 2, or 3 & 4) were drawn from the same dark matter cube. Thus, the expectation is that the set of 500 even numbered mocks and the set of 500 odd numbered mocks will be independent of any correlations caused by the sampling, and any cross correlation will be due to noise. The cross correlation coefficient,

$$\rho_{X,Y} = \frac{\mathsf{C}(X,Y)}{\sigma_X \sigma_Y} \tag{3.37}$$

for both the monopole and quadrupole of the correlation function, and for the power spectrum, calculated from the 500 pairs of mocks drawn from the same dark matter cube is shown in Fig. 3.18. The dashed lines in Fig. 3.18 indicate the maximum and minimum correlation coefficient (at any scale considered) between 500 pairs of independent mocks (i.e. taking pairs where both mocks have even or odd numbers). The fact that the cross correlation between pairs drawn from the same dark matter cube is almost entirely within these bounds indicates that there is no cross correlation above the level of noise in the combined MGS covariance matrix, even on scales where the pairs of mocks could, theoretically, be covariant.

3.6.2 Random Catalogue Redshift Assignment

The effect of assigning redshifts to the random data points from randomly chosen galaxies, as opposed to simply generating them by sampling a smooth fit to the number density, is also tested. Fig. 3.19 presents the differences in the measured correlation function monopole and quadrupole moments of the MGS data, when they are calculated using either random data points that are assigned redshifts from the corresponding galaxy catalogue ('shuffled'), or when they are given redshifts sampled from the fitted number density described in Section 3.1. 'Shuffling' may be expected to reduce the clustering, especially on scales below $100\,h^{-1}\,\mathrm{Mpc}$, because radially averaged features in the galaxy field are removed in the shuffled approach. The power removed is predominantly along the line of sight, and hence the quadrupole is affected more than the monopole. However, Fig. 3.19 shows that for both monopole and quadrupole the difference in clustering

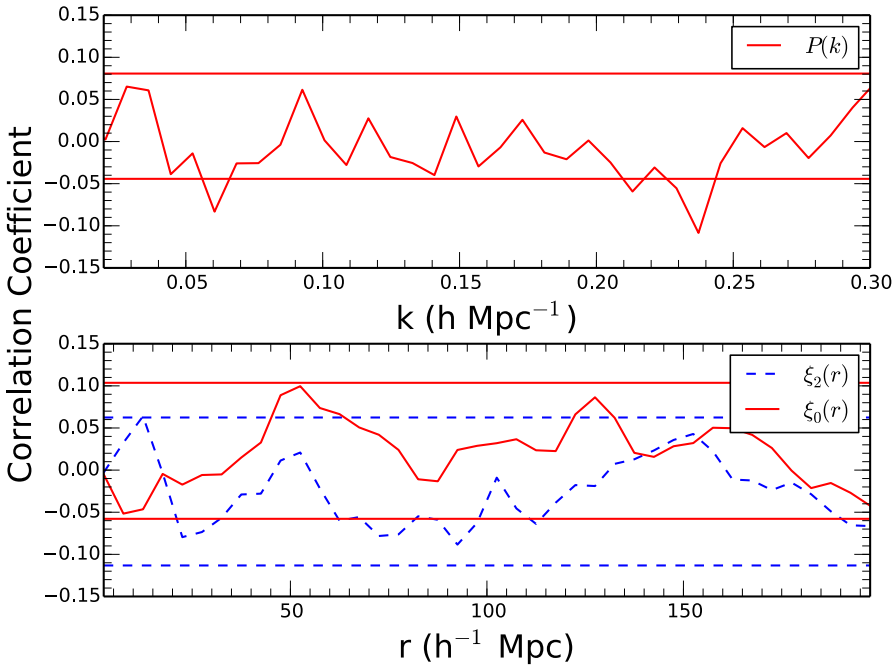


Figure 3.18: The cross-correlation coefficient between pairs of mocks generated from the same dark matter field, for both the power spectrum and the monopole and quadrupole of the correlation function. The horizontal lines indicate the maximum and minimum (across all scales) cross-correlation measured from an equivalent number of pairs of mocks that are drawn from different dark matter realisations.

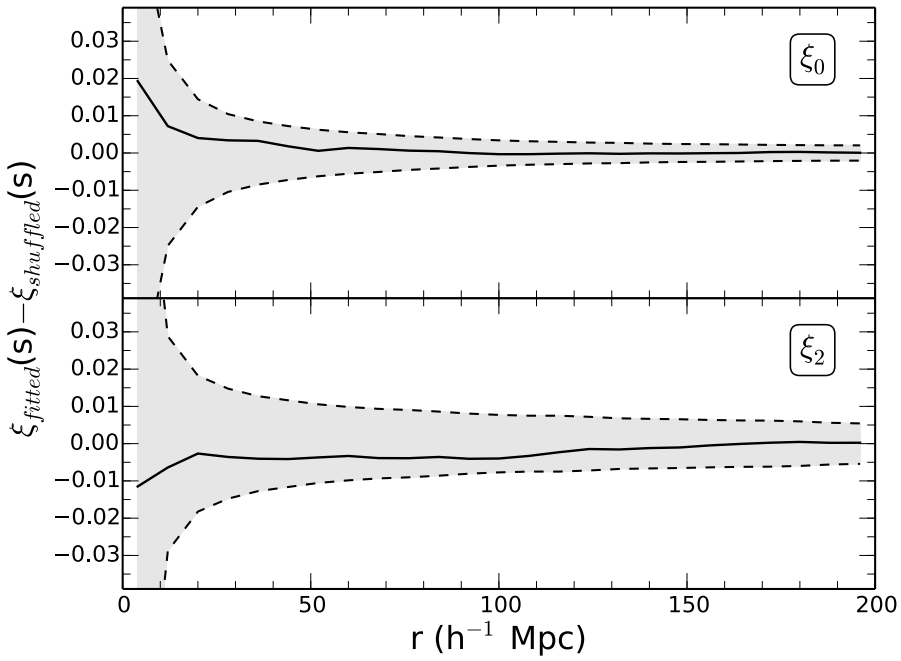


Figure 3.19: The difference in the monopole and quadrupole of the correlation function measured from the data when the fitted and shuffled methods are used to generate redshifts for random data points. The shaded areas denote the one-sigma error regions. The difference between the two methods is well within the one-sigma region on all scales.

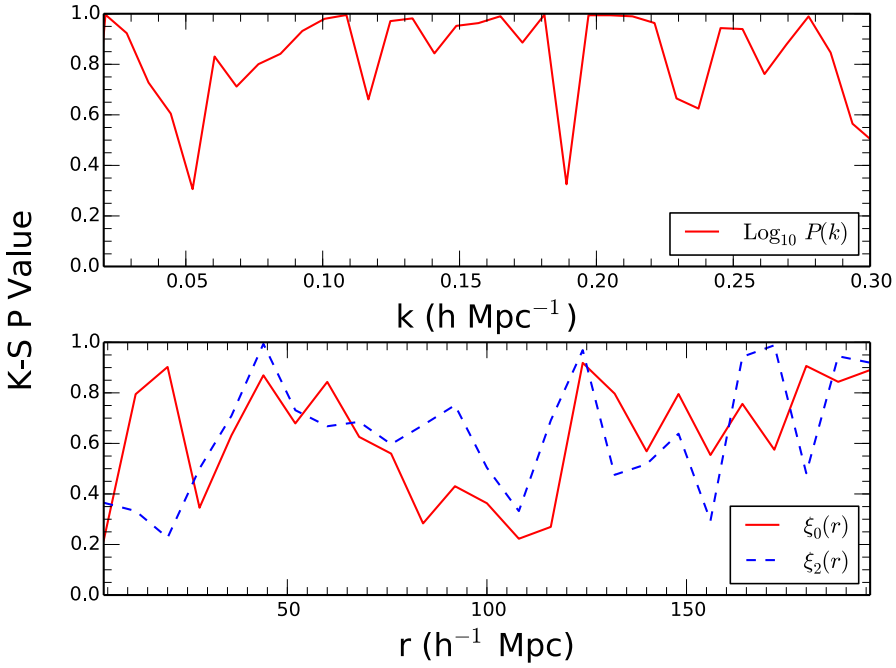


Figure 3.20: The Kolmogorov-Smirnov p-value for both the log of the power spectrum and the monopole and quadrupole of the correlation function. For both statistics the probability that they are drawn from a multivariate Gaussian is high, allowing for computation of the likelihoods for theoretical models from the chi-squared difference between the model and data.

between the two methods is well below the level of the noise. In future analyses the shuffling approach is apdopted as the true radial distribution for the data is not fully known, and this approach accounts for all additional features caused by the galaxy selection, at the expense of a small reduction in the monopole and quadrupole moments. Furthermore, Ross et al. (2012) found that the shuffling approach is less biased than fitting to a smooth $n(z)$ when both methods were tested on BOSS mocks (with a known $n(z)$), and the differences found here are consistent with those of Ross et al. (2012). Such differences are so small that it is not necessary to account for these during model fitting.

3.6.3 Gaussianity of Data

One final test is on the assumption that the measured correlation function and power spectrum are drawn from an underlying multivariate Gaussian distribution. This assumption is often made during BAO and RSD model fitting as it allows the likelihood of any theoretical models given the data to be quantified using the chisquared difference between the two.

A Kolmogorov-Smirnov test is performed on the log of the power spectrum (which

is more commonly used for BAO fitting) and monopole and quadrupole of the MGS mock catalogues, using the cumulative distribution function (CDF) of the normalised differences between the two-point statistics measured from each mock realisation and the average over all the mock catalogues.

Following the standard method of the Kolmogorov-Smirnov test one can define the parameter D as the maximum difference between the measured CDF and the CDF of the distribution one wishes to test against, in this case a Gaussian. The p-value for this test, which indicates the probability that the observed value of D would be as large as it is if our underlying distribution *were* Gaussian, is then given by a simple rescaling of the parameter D,

$$D^* = D\left(\sqrt{N} + \frac{0.11}{\sqrt{N}} + 0.12\right), \tag{3.38}$$

and the approximate expression

$$P(D > D_{obs}) \approx 2\sum_{k=1}^{\infty}(-1)^{k-1}e^{-2k^2D^*}. \tag{3.39}$$

Here, N is the number of bins in the measured CDF. As elsewhere, bins of width $\Delta k = 0.008\,h\,\mathrm{Mpc}^{-1}$ were used for the power spectrum and $\Delta s = 8\,h^{-1}\,\mathrm{Mpc}$ for the correlation function.

Fig. 3.20 shows the Kolmogorov-Smirnov test p-value for the two-point statistics as a function of scale. There is no apparent trend with scale and across all scales of interest the p-value indicates a high probability that both the power spectrum and correlation function are drawn from a Gaussian distribution. The log of the power spectrum has a particularly high probability of being drawn from a Gaussian distribution, which was expected *a priori*, and is why this is often used rather than the power spectrum itself when fitting the BAO feature. Based on the p-values obtained, even for those bins in the correlation function where the difference between the measured CDF and a Gaussian CDF is largest, a greater difference could still be expected at least 20% of the time if the measured clustering statistics were drawn from an underlying Gaussian distribution.

3.7 Summary

In this chapter we have presented the production of a set of 1000 mock catalogues based on a galaxy sample derived from the Sloan Digital Sky Survey Data Release 7 dataset. This sample, dubbed MGS, consists of $63,163$ Luminous Red galaxies over 6813 deg^2 with a low effective redshift of $z_{eff} = 0.15$. Obtaining cosmological constraints from this dataset requires accurate estimates of the covariance matrix and precise knowledge of systematics, which is the driving force behind the creation of the MGS mocks.

This chapter has detailed and verified the steps in the mock creation procedure, including using L-PICOLA to create 500 accurate dark matter fields, the development of an MPI-based parallel, FoF code to identify halos within each simulation, and the population of these halos with galaxies using the HOD method. Masking and subsampling the mocks ensures that these match the window function of the MGS data. The complete set of 1000 mocks matches the clustering of data extremely well, within $\sim 5\%$ down to $k = 0.3\,h\,\mathrm{Mpc}^{-1}$ and on scales as low as $10\,h^{-1}\,\mathrm{Mpc}$.

This chapter has introduced how the MGS mocks can be used to estimate the covariance matrix for the power spectrum and correlation function and perform simple systematic tests on common analysis methods used for large scale structure measurements. In the next chapter the MGS data and mocks will be further utilised to produce robust low redshift BAO and RSD measurements and cosmological constraints. Testing of the fitting models and methods used for this analysis, enabled by the mock catalogues, will be presented therein.

Chapter 4

Measuring the BAO and RSD Signals of the MGS.

This chapter presents measurements and subsequent cosmological interpretation of the BAO and RSD signals within a subset of the Sloan Digital Sky Survey Data Release 7 Main Galaxy Sample detailed in Chapter 3. In this previous chapter, the desire for a set of low-redshift BAO and RSD measurements, providing robust cosmological constraints on the dark energy equation of state, expansion rate of the universe and growth rate of structure, motivated the creation of the MGS dataset. This dataset, consisting of 63,163 highly biased, luminous red galaxies in the redshift range $0.07 < z < 0.2$, is further complemented by a set of 1000 detailed mock galaxy catalogues. The creation of these, along with the presentation of the MGS data, was the main focus of the previous chapter.

This chapter in turn provides a comprehensive view of the BAO and RSD results obtained, with emphasis on the RSD fitting methodology and procedure. These results are tested for robustness and then further used to provide a new set of cosmological constraints. The chapter is laid out as follows: Section 4.1 details the BAO results from the MGS data and provides a brief overview of how these were obtained. Section 4.2 gives an overview of the model used to measure the RSD signal in the MGS data. Section 4.3 uses this model to fit the mocks and performs a series of tests on the accuracy and robustness of the model. Section 4.4 then presents anisotropic RSD fits to the MGS data. Finally, Section 4.5 shows the cosmological constraints on the dark energy equation of state, expansion rate and growth rate from the MGS dataset in comparison to other studies, fulfilling the original motivation for the creation of the MGS.

4.1 Measuring the BAO Scale at $z = 0.15$

The methodology used to measure isotropic BAO positions from the correlation function and power spectrum of the MGS data is adapted from and nearly identical to that used for the BOSS-DR11 BAO analysis (Anderson et al., 2014b; Tojeiro et al., 2014). The quantity of interest in the fit is the dilation parameter α which was previously defined in Eq 1.108. This can be constrained by comparing the BAO peak position in the data to a fiducial model, and in turn can be related to the isotropic combination of the angular diameter distance and Hubble parameter presented in Eq 1.107.

The measured, spherically averaged, monopole correlation function and power spectrum are fit separately and then the results are combined, using the mocks to quantify the correlation coefficient between these two measurements. For the most part, the power spectrum and correlation function contain the same information, the cross-correlation coefficient between these two statistics as measured from the mocks is ~ 0.98. However the methods for fitting these two statistics and the exact range of scales probed when fitting these is slightly different, such that combining the two different statistics improves the statistical power and can remove systematic biases in the fitting method for a single measurement. For fits to both the correlation function and power spectrum, the best-fit values of α are obtained assuming that $\xi(s)$ and $\log P(k)$ were drawn from multi-variate Gaussian distributions and by calculating χ^2 at intervals of $\Delta\alpha = 0.001$ in the range $0.8 < \alpha < 1.2$. This assumption was tested and verified in Chapter 3, where it was found to be a good approximation, especially for $\log P(k)$.

The data and mocks are fit both pre- and post-'reconstruction', a process which serves to linearise and sharpen the BAO feature, improving the measurements of the acoustic scale.

4.1.1 Reconstruction

One of the key limitations on the statistical power of the BAO feature for cosmological constraints is the effect of gravitational evolution. The evolution of the galaxy positions, which on average causes displacements of $\sim 10\,h^{-1}\,\mathrm{Mpc}$ (Burden et al., 2014), distorts the BAO signal and degrades the precision with which the characteristic scale of the feature can be measured, broadening the BAO peak in configuration space and damping the oscillations in the power spectrum. This is in addition to similar effects caused by Redshift Space Distortions.

To improve the precision on the measurements of the BAO scale from the MGS data, an algorithm, designed to 'reconstruct' the linear density field based on the measured non-linear overdensity field, was applied. The reconstruction method uses the measured

galaxy map to construct a displacement field that is used to redistribute the galaxies into a spatial configuration that more closely reproduces their initial positions and removes the effect of redshift space distortions. This process thereby (typically) sharpens the BAO feature in clustering measurements and removes non-linear shifts in its peak position, thus allowing significantly more precise and accurate BAO measurements. Reconstruction of galaxy clustering data (Eisenstein et al., 2007b) has been shown to improve measurements of the BAO scale in multiple galaxy samples, including SDSS-II LRGs at $z = 0.35$ (Padmanabhan et al. 2012 and Xu et al. 2013), the SDSS-III BOSS LOWZ and CMASS samples (Anderson et al. 2014b and Tojeiro et al. 2014), red and blue galaxies in the CMASS sample (Ross et al., 2014) and emission line galaxies from the WiggleZ survey (Kazin et al., 2014).

For the MGS, fixed values of $b = 1.5$ and $f = 0.6413$ and a smoothing scale of $15\,h^{-1}\,\mathrm{Mpc}$ are adopted. Although the exact values of these parameters are somewhat arbitrary and unknown *a priori* (indeed the aim of performing RSD fits to the MGS data is to constrain f and b), Anderson et al. (2014b) and Burden et al. (2014) find that the reconstruction method is relatively insensitive to small changes ($\sim 20\%$) in these values. I.e., using simulations with a known bias and growth rate, reconstructing the galaxies using the true, correct values and values that are 'incorrect' by up to 20% gives comparable results.

The clustering of the galaxy sample, pre-(grey diamonds) and post-reconstruction (open circles), is displayed in Fig. 4.1 for the correlation function and in Fig. 4.2 for the power spectrum. One can see, most easily by studying the $P(k)$ measurements, that reconstruction induces a decrease in the clustering amplitude that is nearly constant (at scales less than the BAO scale for $\xi(s)$). This is due to the removal of large-scale redshift-space distortions. One can further see, most easily by studying the $\xi(s)$ measurements, that reconstruction sharpens the BAO feature. Throughout the rest of the section, attention is focused on BAO measurements obtained from the post-reconstruction measurements.

4.1.2 Fitting the BAO Signal in the MGS Mocks

The BAO fitting methodology described above is tested by measuring the BAO scale of the mock galaxy samples. This also allows for characterisation of the expected constraints from the fits to the MGS data. All results are quoted as the best-fit value with 1 σ defined as the $\Delta\chi^2 = 1$ region. Fits are performed on both the individual mocks and the mean of the mocks. The former set of fits allows for investigation into the range of α values one can expect to recover and whether the fit to the data, and the measured BAO signal itself, is typical of what would be expected. The quoted fits in this section are performed on the mean of the mock samples which allows for a sensitive test of systematics in the

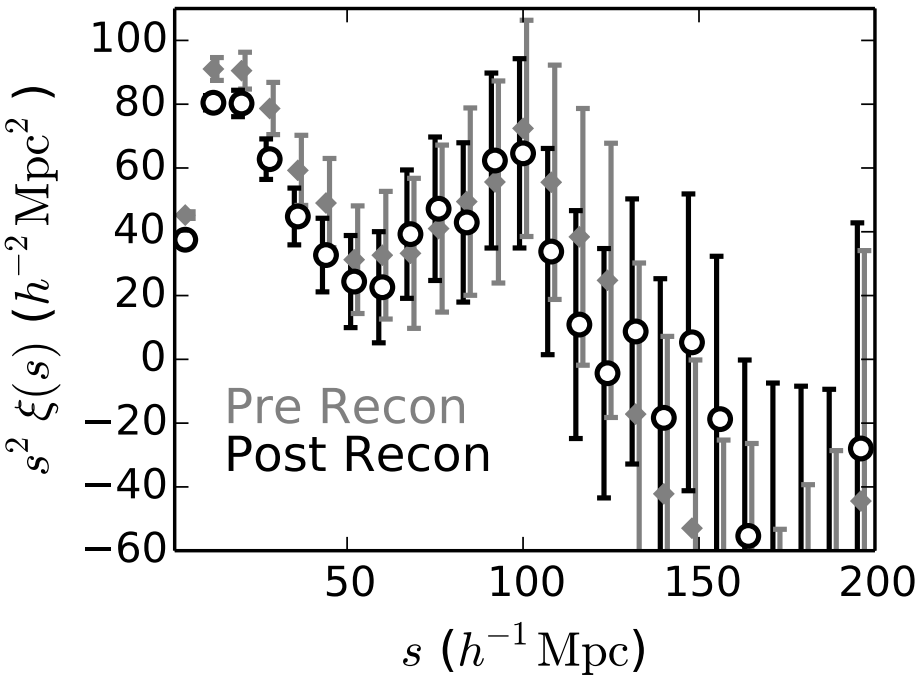


Figure 4.1: The measured correlation function, $\xi(s)$ (points with error-bars), pre- (grey diamonds) and post- (open circles) reconstruction. The error-bars are determined from the variance of the 1000 mock galaxy samples. One can see that reconstruction reduces the clustering amplitude, due to removal of large-scale redshift space distortions, and sharpens the BAO peak.

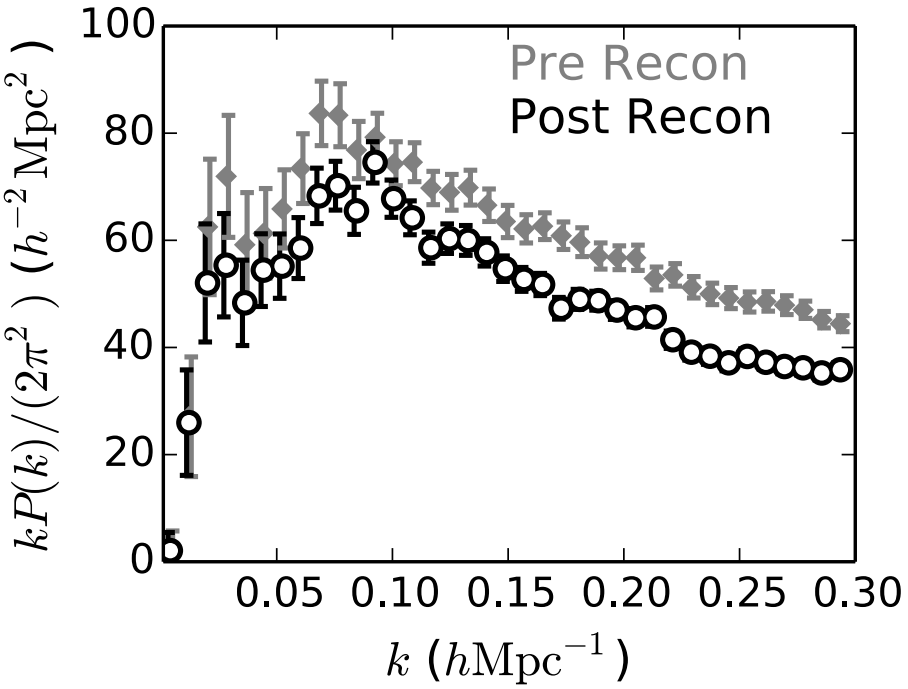


Figure 4.2: The measured power spectrum, $P(k)$ (points with error-bars), pre- (grey diamonds) and post- (black circles) reconstruction. The error-bars are determined from the variance of the 1000 mock galaxy samples. The clustering amplitude of the post-reconstruction data is decreased across all k due to the removal of large-scale redshift space distortions.

fitting method. Using the mean of the mocks produces a smooth BAO feature without the noise present in a single realisation. In essence the fits should then be able to recover the fiducial cosmology very well, and any differences between the measured and fiducial cosmology are the result of systematics. The magnitude of any difference when fitting to the mean of the mocks can be compared to the expected error on a single realisation to check that the systematic errors are negligible in comparison with the statistical errors in the data. The fiducial cosmology used for the fits is the same as that used to generate the mocks and as such the expected best fit value is $\alpha = 1.0$.

Some of the MGS mocks result in a very low detection of the BAO. Additionally, a small number of our mocks have a reasonable BAO detection but result in α values that are far from the mean. These extreme cases tend to correspond to low values of α and artificially skew the recovered distribution. The reason for this is that low values of α push the best-fitting model towards large scales, dominated by low numbers of pair counts and the survey window, where the clustering becomes excessively noisy. This noise in turn makes it 'easy' to find a particular BAO feature which fits well, but which corresponds to cosmological models well outside current constraints from other probes and surveys. To avoid these scenarios only numbers for mocks that have a 2σ bound, and hence solid detections, are quoted, and $> 4\sigma$ outliers are excluded.

Averaging over the mocks, values of $\alpha = 0.996 \pm 0.002$ for $P(k)$ and $\alpha = 0.997 \pm 0.002$ for $\xi(s)$ are found. Averaging these two results, taking into the high cross-correlation between the two statistics, gives $\alpha = 0.996 \pm 0.002$. The average is taken as the exact range of scales probed by these statistics differs slightly and hence the information inherent in their combination is larger than in a single measurement. This result is more than 2σ away from 1 (1σ calculated via $0.047/\sqrt{895}$ where the denominator is the number of mocks that are retained). This result is partially driven by outliers as excluding fifteen additional $> 3\sigma$ outliers, increases the mean to 0.997. However, some small bias remains. The magnitude of this bias is less than 0.1 of the 1σ expected uncertainty in the measurement from the data, and thus not significant in any application of the measurements. Given that the same methodology was used as Anderson et al. (2014b), who found no detectable bias in isotropic BAO measurements, it is not believed that this bias is suggestive of a systematic bias in the BAO fitting methodology. Instead, as the MGS data is of much lower quality than the data used in Anderson et al. (2014b), this is believed to be due to the non-Gaussianity present in the fitting of noisy data, which skews the likelihoods recovered from the fit.

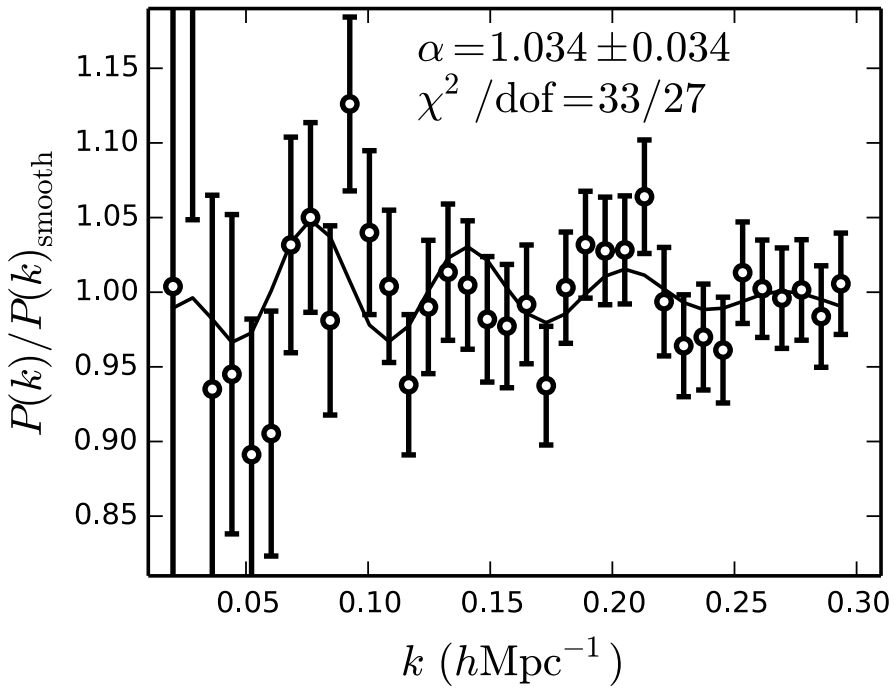


Figure 4.3: The measured post-reconstruction power spectrum, $P(k)$, (points with error-bars) and best-fit model (solid curve) divided by the smooth (no BAO) component of the best-fit model.

4.1.3 BAO Fits to the MGS Data

Detections of the BAO signal are obtained for the reconstructed MGS data using both $P(k)$ and $\xi(s)$. The best-fit $P(k)$ BAO model (divided by the smooth component of the best-fit $P(k)$ model), using the fiducial binning scheme and fitting methodology, is displayed in Fig. 4.3. A value of $\alpha = 1.034 \pm 0.034$ is recovered from $P(k)$ and the best-fit model is a good fit (χ^2/dof = 32.6/27, a greater χ^2 would be expected in 21 per cent of cases). The best-fit model found for the fiducial $\xi(s)$ binning is displayed against the measurement in Fig. 4.4. Here $\alpha = 1.058 \pm 0.036$ is the best fit value. Similar to $P(k)$, the χ^2/dof is slightly greater than one (20.3/16) and represents a good fit (again a greater χ^2 would be expected in 21 per cent of cases).

The above results are found to be robust to changes in the fitting procedure, including the scales used for the fit, the binning scheme and the number of terms used to marginalise over the broadband signal. The final $\xi(s)$ and $P(k)$ results are then obtained by combining the results when using different bin widths, scales and fitting procedures, in a way that makes use of the correlations between these measurements. As these measurements are of the same data they are very tightly correlated. These correlations are taken into account by using the mock catalogues to estimate the cross-correlation between the dif-

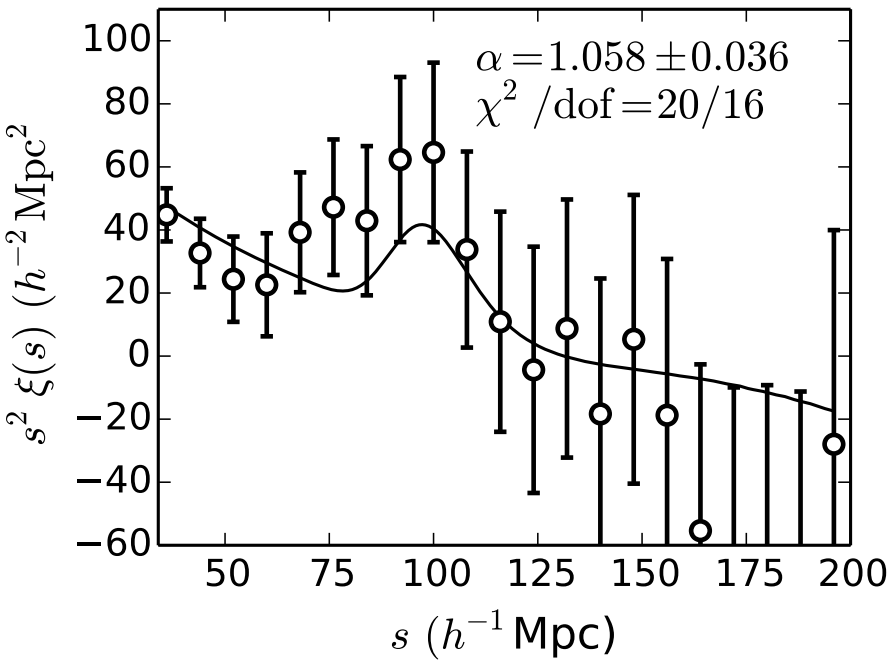


Figure 4.4: The measured post-reconstruction correlation function, $\xi(s)$, (points with error-bars) and best-fit BAO model (solid curve) .

ferent measurements of the two statistics. As the cross-correlation is very high (there is little extra information in the different measurements of the same statistic, or between the $\xi(s)$ and $P(k)$ measurements) the change in the value and error when combining the measurements is small, however the final results are more robust than that from a single measurement as systematic effects that could arise when fitting to a single range of scales or with a single bin width are reduced.

The final $\xi(s)$ (1.050 ± 0.040) measurement is less precise than its $P(k)$ counterpart (1.031 ± 0.034). The difference in the uncertainties is typical of what was found in the mocks samples and the difference between the best-fit value of $P(k)$ and $\xi(s)$ is of similar significance to that found in Anderson et al. (2014b). The final consensus measurement of $\alpha = 1.040 \pm 0.037$ is obtained by combining these two measurements, again accounting for the cross-correlation between the two statistics estimated from the mocks.

The differences between the $P(k)$ and $\xi(s)$ BAO measurements are large for the data compared to the mocks, though a similar result is found in Anderson et al. (2014b). The precision of the MGS measurements is a factor of four less than that of Anderson et al. (2014b), suggesting that a common explanation requires greater scatter between the results obtained from mock samples (rather than a bias in one particular measurement technique). It is conceivable that including more realism (e.g., light-cones) in mock samples would reduce the correlation between the $P(k)$ and $\xi(s)$ BAO measurements recovered

from such mocks, which would explain both the findings of the MGS BAO fits and those of Anderson et al. (2014b). The robustness checks preformed reveal no potential systematic effect that would bias either the $\xi(s)$ and $P(k)$ measurements.

4.2 Modelling the Redshift Space Monopole and Quadrupole

To model the redshift space monopole and quadrupole the combined Gaussian Streaming/Convolved Lagrangian Perturbation Theory (CLPT) model of Wang et al. (2014) is used. The clustering of galaxies in redshift space can be written as a function of their real space correlation and their full pairwise velocity dispersion (Fisher, 1995; Scoccimarro, 2004). In the Gaussian Streaming model, introduced by Reid & White (2011), the pairwise velocity dispersion is approximated as a Gaussian, which allows one to write the two-dimensional redshift space correlation function, $\xi(s_\perp, s_{||})$, as a function of the real-space correlation function, $\xi(r)$, and the mean infall velocity and velocity dispersions betweens pairs of galaxies, $v_{12}(r)$ and $\sigma^2_{12}(r,\mu)$ respectively,

$$1+\xi(s_\perp, s_{||}) = \int_{-\infty}^{\infty} \frac{dr_{||}}{[2\pi\sigma^2_{12}(r,\mu)]^{1/2}}[1+\xi(r)] \times \exp\left\{-\frac{[s_{||}-r_{||}-\mu v_{12}(r)]^2}{2\sigma^2_{12}(r,\mu)}\right\}. \tag{4.1}$$

Here $s_\perp = r_\perp$ and $s_{||}$ denote redshift space separations transverse and parallel to the line of sight, $r_{||}$ denotes the real space separation parallel to the line of sight, such that $r^2 = r_\perp^2 + r_{||}^2$, and $\mu = r_{||}/r$ is as defined previously.

Reid & White (2011) evaluate $v_{12}(r)$ and $\sigma^2_{12}(r,\mu)$ using a standard perturbation theory expansion of a linearly biased tracer density field. Although this provides a good model on large scales, it does not accurately replicate the velocity statistics of the tracer field on small scales, nor the smoothing of the BAO feature. This was improved upon by Reid et al. (2012) in their analysis of the BOSS CMASS galaxy sample by using Lagrangian Perturbation Theory to generate the real-space correlation function above scales of $70\,h^{-1}\,\mathrm{Mpc}$. This proved effective for the BOSS CMASS sample, although Reid et al. (2012) note that the BOSS CMASS galaxy sample has a second order bias close to zero, the point at which the accuracy of the standard perturbation theory evaluation of $v_{12}(r)$ and its derivative is greatest.

Carlson et al. (2013) and Wang et al. (2014) further improve the modelling of the correlation function by computing the real-space correlation function using Convolved Lagrangian Perturbation Theory and evaluating $v_{12}(r)$ and $\sigma^2_{12}(r,\mu)$ in the same framework. This formulation relies on a perturbative expansion of the Lagrangian overdensity

and displacement which in turn allows the correlation function and velocity statistics to be written as a series of integrals over powers of the linear power spectrum. For biased tracers the model assumes a local real-space Lagrangian bias function, F, and solutions up to $\mathcal{O}(P_L^2)$ reveal a dependence on both the first and second derivatives of the bias function, $\langle F' \rangle$ and $\langle F'' \rangle$, and combinations thereof. Furthermore, as would be expected, the velocity statistics have a dependency on the growth rate of structure, f, via the multiplicative factor, f^2. From Matsubara (2008) the linear galaxy bias, b, can be easily related to the first derivative of the Lagrangian bias function via $\langle F' \rangle = b - 1$.

The model is calculated as follows. For a vector $\boldsymbol{r}$ in real space and vector $\boldsymbol{q}$ in Lagrangian space, three functions are defined, which depend on the Lagrangian bias, growth rate and linear power spectrum: $M_0(\boldsymbol{r}, \boldsymbol{q}, \langle F' \rangle, \langle F'' \rangle, f, P_L)$,
$M_{1,n}(\boldsymbol{r}, \boldsymbol{q}, \langle F' \rangle, \langle F'' \rangle, f, P_L)$ and $M_{2,nm}(\boldsymbol{r}, \boldsymbol{q}, \langle F' \rangle, \langle F'' \rangle, f, P_L)$. M_0 is a scalar function, whilst $M_{1,n}$ and $M_{2,nm}$ are vector and tensor functions along cartesian directions n and m. The exact form of the functions M_0, $M_{1,n}$, and $M_{2,nm}$ are given in Wang et al. (2014).

$\xi(r)$ and $v_{12}(r)$ are calculated by projecting the scalar and vector functions along the pair separation vector and integrating with respect to the Lagrangian separation,

$$1 + \xi(r) = \int d^3q M_0(\boldsymbol{r}, \boldsymbol{q}), \tag{4.2}$$

$$v_{12}(r) = [1 + \xi(r)]^{-1} \int d^3q M_{1,n}(\boldsymbol{r}, \boldsymbol{q}) \hat{r}_n. \tag{4.3}$$

The velocity dispersion $\sigma_{12}^2(r, \mu)$ is split into components perpendicular and parallel to the pair separation vector. These are evaluated separately by projecting and integrating the tensor function,

$$\sigma_{12}^2(r, \mu) = \mu^2 \sigma_{||}^2(r) + (1 - \mu^2)\sigma_{\perp}^2(r), \tag{4.4}$$

where

$$\sigma_{||}^2(r) = [1 + \xi(r)]^{-1} \int d^3q M_{2,nm}(\boldsymbol{r}, \boldsymbol{q}) \hat{r}_n \hat{r}_m, \tag{4.5}$$

$$\sigma_{\perp}^2(r) = \frac{[1 + \xi(r)]^{-1}}{2} \int d^3q M_{2,nm}(\boldsymbol{r}, \boldsymbol{q}) \delta_{nm}^K - \frac{\sigma_{||}^2}{2} \tag{4.6}$$

and δ_{nm}^K is the Kronecker delta.

Hence, for a given cosmological model parametrised by $P_L, b, \langle F'' \rangle$ and f, the CLPT model can calculate, for any scale of interest, a unique set of $\xi(r)$, $v_{12}(r)$ and $\sigma_{12}^2(r, \mu)$. Entering these into Eq. (4.1) generates the two-dimensional redshift space correlation function and from there a model monopole and quadrupole can be obtained in the standard way. These models are fitted to the measurements from data and mocks as described later to constrain a given set of cosmological parameters.

4.2.1 Alcock-Paczynski Effect

In calculating the correlation function, a (fiducial) cosmological model must be adopted to calculate the physical separations between galaxies parallel and transverse to the line of sight. Specifically, to calculate the separation along the line of sight requires the Hubble parameter, $H(z)$, and the galaxy redshifts, whilst the transverse separation requires knowledge of the angular diameter distance, $D_A(z)$, and the angular separation of the galaxy pair. Any difference between the relative values of these parameters in the fiducial cosmology and the true cosmology will manifest as anisotropic clustering, that is, a difference in the clustering of galaxies parallel and perpendicular to the line of sight. If an observable, such as the BAO feature, is expected to be statistically isotropic, then any measured anisotropy can also be used to constrain the true cosmology of the Universe. This is the Alcock-Paczynski test (AP; Alcock & Paczynski 1979). This effect is in addition to the anisotropy added by Redshift Space Distortions. As such, the AP effect and RSD are degenerate and a way to disentangle these effects is needed.

Following Xu et al. (2013), two scale parameters are introduced, α and ϵ. α has been previously introduced for the BAO fitting method and denotes the stretching of all scales and encapsulates the isotropic shift whilst ϵ parametrises the AP effect. Measuring these two parameters allows one to constrain the angular diameter distance and Hubble expansion independently,

$$\alpha = \left(\frac{D_A^2(z)}{D_{A,fid}^2(z)} \frac{H_{fid}(z)}{H(z)} \right)^{1/3} \frac{r_{s,fid}}{r_s}, \tag{4.7}$$

$$1 + \epsilon = \frac{F(z)}{F_{fid}(z)} = \left(\frac{D_{A,fid}(z)}{D_A(z)} \frac{H_{fid}(z)}{H(z)} \right)^{1/3}. \tag{4.8}$$

where a subscript 'fid' denotes the fiducial model. Values $\alpha = 1.0$ and $\epsilon = 0.0$ would indicate that the fiducial cosmology is the true cosmology of the measured correlation function.

For a given model correlation function the α and ϵ parameters modify the scales at which a given value is measured for the correlation function,

$$\begin{aligned} s'_{||} &= \alpha(1+\epsilon)^2 s_{||}, \\ s'_{\perp} &= \alpha(1+\epsilon)^{-1} s_{\perp}. \end{aligned} \tag{4.9}$$

Values of α and ϵ are applied directly to the model, altering the scales at which the two-dimensional redshift space correlation function (given by Eq. (4.1)) is calculated. The necessary corrections to the parallel and perpendicular separations, $s_{||}$ and $s_{\perp}$, are evaluated before these are used to calculate the corresponding values of $r, r_{||}$ and μ required by the integrand. Subsequently the 2D model for the correlation function is integrated to estimate the monopole and quadrupole moments.

4.2.2 Correction for Binning Effects

Finally, the way the data is binned must be accounted for when calculating the model correlation function. Rather than evaluating the model at the centre of the bins, variations in the model correlation function across each bin are included by instead taking the weighted average of the model within each bin. For a bin from s_1 to s_2 centred at s, the model is

$$\xi_{0,\mathrm{mod}}(s) = \frac{1}{V}\int_{s_1}^{s_2} \xi_0(s')s'^2 ds',$$
$$\xi_{2,\mathrm{mod}}(s) = \frac{1}{V}\int_{s_1}^{s_2} \xi_2(s')s'^2 ds'. \tag{4.10}$$

Where V is the normalisation for the weighted mean,

$$V = \int_{s_1}^{s_2} s'^2 ds'. \tag{4.11}$$

For all the RSD fits detailed in this chapter the model is evaluated in bins of width $\Delta s = 1\,h^{-1}\,\mathrm{Mpc}$ between $0\,h^{-1}\,\mathrm{Mpc} < s \leq 200\,h^{-1}\,\mathrm{Mpc}$. Eq. (4.10) is then calculated using a cubic spline interpolation method to interpolate the value of the monopole and quadrupole at any point required for the integration.

4.2.3 Cosmological Parameters

For the RSD analysis, the shape of the linear power spectrum is considered to be parametrised by the cold dark matter and baryonic matter densities, $\Omega_c h^2$ and $\Omega_b h^2$, and the scalar index, n_s, whilst the amplitude of the power spectrum is quantified using σ_8. Additionally, there are the growth rate of structure, f, which will be measured via the RSD signal, galaxy bias parameters b and $\langle F'' \rangle$, and BAO dilation parameters α and ϵ. These latter two parameters are measured independently of the power spectrum shape.

In theory, the dependence of the CLPT model on P_L, b, f, $\langle F'' \rangle$ and σ_8 is such that, combined with the other dependencies, all of the above parameters can be independently measured if the data has no noise. In practice however, the parameters f, b and σ_8 are strongly degenerate and cannot be constrained independently. In addition, no constraints on the shape of the linear power spectrum can be obtained beyond those, already tight, constraints given by the Planck Collaboration's analysis of the Cosmic Microwave Background radiation. In lieu of this $\Omega_c h^2$, $\Omega_b h^2$ and n_s are fixed to the fiducial values used to create the MGS mock catalogues, which correspond closely to the Planck best-fit values. The assumption is then that any variation in these parameters can be captured by departures from $\alpha = 1.00$ and $\epsilon = 0.00$.

Overall, the combination of cosmological parameters to be explored is $\vec{p} = \{b\sigma_8, \langle F'' \rangle, f\sigma_8, \sigma_{8,nl}, \alpha, \epsilon\}$. Here σ_8 is treated as containing two separate contributions, linear and

non-linear. The former of these is contained in the $b\sigma_8$ and $f\sigma_8$ parameters which are the parameters of interest and are responsible for the overall amplitude of the monopole and quadrupole of the correlation function. The latter, $\sigma_{8,nl}$, is only effective at the smallest scales that are fit and as such is largely unconstrained and degenerate with the second order bias parameter $\langle F'' \rangle$.

In all the RSD fits $f\sigma_8$ is not allowed to vary in such a way that unphysical values of $f\sigma_8 < 0$ or $\sigma_{8,nl} < 0\,h^3\,\mathrm{Mpc}^{-3}$ are chosen. Uniform priors of $0.8 < \alpha < 1.2$ and $-0.2 < \epsilon < 0.2$ are applied as for the BAO fits. Additional priors on α and $\sigma_{8,nl}$, based on cosmological results from other probes and surveys, are described and tested in Section 4.3 and applied to the final results where appropriate.

4.2.4 Nuisance Parameters

There are two nuisance parameters which are marginalised over while fitting the correlation function, denoted σ_{offset} and IC. The first of these corresponds to an additive correction to σ_{12} in the Gaussian Streaming model. This compensates for two different effects that both manifest at the same point in the model. The first is the CLPT model's inability to fully recover the large scale halo velocity dispersion. Whilst the scale-dependence of both the $\sigma_{||}$ and $\sigma_{\perp}$ parts of σ_{12} is well recovered by the CLPT, there is a mass-dependent, constant amplitude shift across all scales. This systematic offset in the halo velocity dispersion offset is identified in Reid & White (2011) and further explored in Wang et al. (2014), who go on to suggest that it stems from gravitational evolution on the smallest scales, which cannot be accurately predicted by perturbation theory and hence cannot be separated from the overall scale-dependence of σ_{12}. Rather than calibrate the corrective factor required to shift the amplitude of the velocity dispersion using, i.e., N-Body simulations, this is simply treated as a free parameter, and part of the σ_{offset} nuisance parameter. The second component of σ_{offset} is the additional velocity dispersion along the line of sight due to the so called, 'Fingers-of-God', resulting from peculiar motions of the galaxies within their host halos. This effect is expected to be small on the scales of interest to the RSD fit in the monopole and quadrupole of the correlation function.

A very broad, flat prior of $-40\,\mathrm{Mpc}^2 < \sigma_{offset} < 40\,\mathrm{Mpc}^2$ is applied. This range is similar to that used in Reid et al. (2012), where they allow the Fingers-of-God intra-halo velocity dispersion to vary from $0\,\mathrm{Mpc}^2$ to $40\,\mathrm{Mpc}^2$, providing a detailed set of tests to validate this prior. In the MGS fits, this is additionally allowed to go negative over the same range to account for the fact that, as mentioned in Reid & White (2011), the perturbation theory calculation of σ_{12} overestimates the amplitude of the positive offset required to bring linear theory into line with the measurements from N-Body simulations,

hence resulting in a σ_{12} which is larger than would be measured.

The second nuisance term is the integral constraint, which takes the form of an additional constant added to the correlation function monopole. This accounts for incorrect clustering on the largest scales due to the finite volume of our survey. Whilst, given a model, this can be calculated analytically from the properties of the survey, it is included as a free parameter to also account for additional uncertainties in the modelling of the monopole and potential observational systematic effects, which tend to add nearly scale-independent clustering (Ross et al., 2012). Under the assumption that the integral constraint is independent of the angle to the LOS, this vanishes for the quadrupole and so a nuisance parameter of this form is applied only to the monopole.

4.3 Fitting the RSD Signal in the MGS Mocks

As with the BAO fits, the RSD model and fitting methodology are tested by fitting the average monopole and quadrupole of the correlation function recovered from the 1000 mocks. The joint covariance matrix appropriate for a single realisation is used, including the cross-covariance between the monopole and quadrupole: thus the errors recovered should match those from a single realisation. Using the average of the mocks results in a smooth correlation function without any of the noise present in a single realisation and so allows for an accurate test of any systematic effects in the modelling. In the absence of systematics and the noise on a given realisation the theory should match the measurements exactly. Using the covariance matrix from a single realisation then means that any systematic differences between the theory and smooth correlation function measured from the mean of the mocks can be quantified in terms of the statistical error on the data, i.e., one can ensure that any systematic effects account for only a small percentage of the total error budget. To fit the clustering, an MCMC sampling over models is performed using the publicly available EMCEE python routine (Foreman-Mackey et al., 2013). For each parameter, the best-fit value of the marginalised likelihood, with 1σ errors defined by the $\Delta\chi^2 = 1$ region around this point, are quoted.

The fiducial fitting choices are as follows: $\Delta s = 8\,h^{-1}\,\mathrm{Mpc}$ is the fiducial bin width, and only those bins with centres $25\,h^{-1}\,\mathrm{Mpc} \leq s \leq 160\,h^{-1}\,\mathrm{Mpc}$ are used. A prior on α based on the results of the independent BAO fits detailed in this chapter is used, as well as priors on ϵ and $\sigma_{8,nl}$ based on results using data from the Planck satellite (Planck Collaboration et al., 2014b). The fiducial range of scales is chosen based on the facts that including larger scales adds little extra information and the accuracy of the CLPT model starts to decrease below $s = 25\,h^{-1}\,\mathrm{Mpc}$ for the range of halo masses where galaxies in the MGS are found (Wang et al., 2014). Other choices are motivated where appropriate in this section and in all cases it is demonstrated that the $f\sigma_8$ measurements are largely

Table 4.1: The mean values and one-sigma errors on $f\sigma_8$ and $b\sigma_8$ from the average of the mocks, recovered from the marginalised probability distribution when different priors are applied (cases 1, 2 and 3), the range of scales and bin sizes used in the fits are changed (cases 4, 5, and 6), when the BAO peak position in the monopole/quadropole is fixed rather than marginalised over (cases 7, 8, and 9) and when a linear model is used (case 10). The expected values for the mocks are $f\sigma_8 = 0.466$ and $1.15 \leq b\sigma_8 \leq 1.22$.

	Average of Mocks:		
No.	Case	$f\sigma_8$	$b\sigma_8$
1	Full fit	$0.43^{+0.47}_{-0.32}$	$1.04^{+0.19}_{-0.18}$
2	Prior on α	$0.49^{+0.28}_{-0.29}$	$1.09^{+0.14}_{-0.19}$
3	Prior on $\sigma_{8,nl}$	$0.45^{+0.19}_{-0.23}$	$1.19^{+0.12}_{-0.13}$
4	$35 \leq s \leq 140\,h^{-1}\,\mathrm{Mpc}$	$0.50^{+0.23}_{-0.24}$	$1.16^{+0.16}_{-0.18}$
5	$\Delta s = 5\,h^{-1}\,\mathrm{Mpc}$	$0.45^{+0.18}_{-0.22}$	$1.20^{+0.11}_{-0.13}$
6	$\Delta s = 10\,h^{-1}\,\mathrm{Mpc}$	$0.42^{+0.17}_{-0.20}$	$1.20^{+0.10}_{-0.11}$
7	$\epsilon = 0.00$	$0.50^{+0.13}_{-0.12}$	$1.18^{+0.10}_{-0.10}$
8	$\alpha = 1.00, \epsilon = 0.00$	$0.50^{+0.13}_{-0.12}$	$1.18^{+0.08}_{-0.08}$
9	$\alpha = 1.04, \epsilon = 0.00$	$0.52^{+0.13}_{-0.12}$	$1.24^{+0.08}_{-0.09}$
10	Linear Fit	$0.42^{+0.11}_{-0.11}$	$1.14^{+0.08}_{-0.08}$

independent of these choices.

As an overview, before detailing individual fits, the best fit values for all the fitting cases are collated in Table 4.1. Fig. 4.5 shows the best-fit values for the cases listed in the table along with the ΛCDM prediction of $f\sigma_8$, which closely matches that used in the production of the mock catalogues, and the expected galaxy bias assuming linear theory (Hamilton, 1992). For the fiducial ΛCDM cosmology, and assuming GR, the expected values are $f(z_{eff}) = \Omega_m(z_{eff})^{0.55} = 0.609$ and $\sigma_8(z_{eff}) = 0.766$, and from the HOD fits to the MGS it is expected that $1.5 \leq b \leq 1.6$ depending on the exact scales used to estimate the linear galaxy bias.

In Fig. 4.6 the best-fitting model monopole and quadrupole for the fiducial fit is plotted alongside that measured from the average of the mocks. It can be seen that the CLPT model does remarkably well in modelling the monopole and quadrupole across all the scales that are fit, with only small inaccuracies on the smallest scales and around $s = 100\,h^{-1}\,\mathrm{Mpc}$. The inaccuracies are clearly well below the expected level of noise in

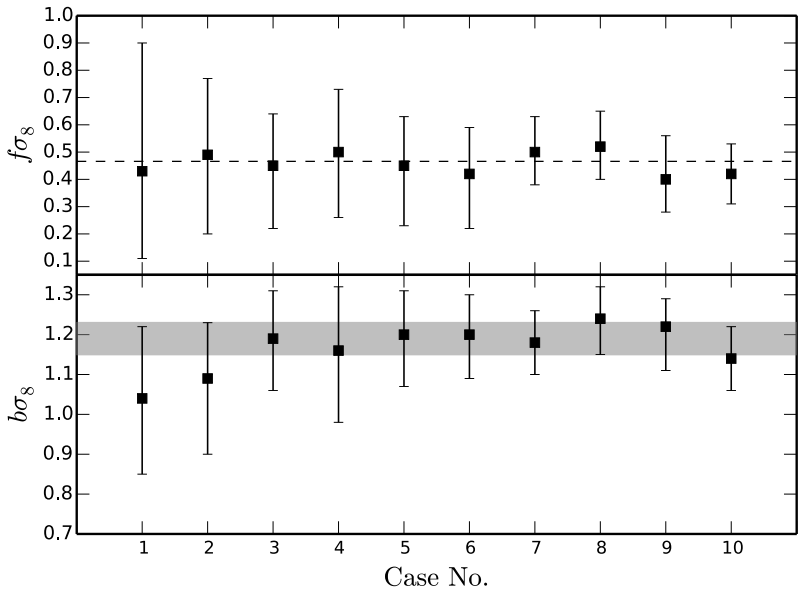


Figure 4.5: The marginalised $f\sigma_8$ and $b\sigma_8$ values and one-sigma errors from fitting to the mean of the mocks for the 10 cases listed in Table 4.1. The dashed line indicates the expected growth rate assuming the fiducial ΛCDM cosmology. The shaded band indicates the expected linear galaxy bias as measured from the HOD fits to the MGS sample, where a band is used rather than a line to account for the fact that the calculated value depends slightly on the range of scales used.

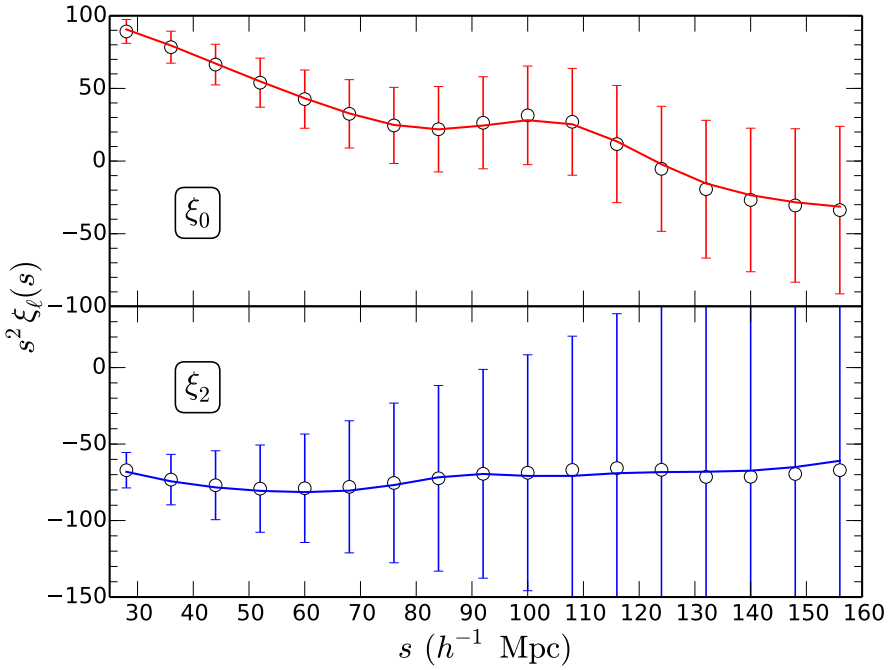


Figure 4.6: The average monopole and quadrupole of the 1000 mock catalogues (points) shown alongside the best-fit model for the fiducial fitting case (solid) which includes both priors on α and $\sigma_{8,nl}$. The errors are derived from the covariance matrix and are the errors on a single realisation. The CLPT model does a fantastic job of reproducing the measured clustering on all scales of interest.

the measurements.

4.3.1 Effects of α Prior

A prior on α is included, motivated by the expected improvement in the BAO peak position after reconstruction, in the fiducial $f\sigma_8$ measurements. The effect of including this for mock results is tested in this section. Much of the information on α comes from the BAO feature. However, in the pre-reconstruction data used to fit the RSD signal the BAO feature in the monopole is very noisy. Reconstruction provides a means to recover more of the information within the BAO feature and hence can improve the constraints on α. During reconstruction a linear RSD model is used to convert the galaxy overdensity in redshift space to a Lagrangian displacement for each galaxy. It is common, but not necessary, to also scale the displacements to remove the linear RSD and simplify the BAO constraints by making the amplitude of the signal isotropic when analysed in the true cosmology. The effect of this process on the quadrupole of the correlation function is not well understood and so post-reconstruction measurements cannot currently be used for RSD constraints.

However, as a result of the independent BAO fits, there is, available, a greater knowledge of α than is apparent in the pre-reconstruction monopole. This is encapsulated using a Gaussian prior on α, centred on the recovered post-reconstruction best-fit values from the BAO fits, and with a variance calculated from the difference between pre- and post-reconstruction fits to the BAO feature (the pre-reconstruction uncertainty is a factor 2.5 times greater than the post-reconstruction result). In other words, it is expected that the inclusion of the α prior will recover the same uncertainty on α as found in the BAO only fitting results. Reconstruction also shifts the position of the BAO peak due to the removal of coupling between different k-modes on the scale of the BAO feature. The BAO fits are performed using the post-reconstruction (hence no mode-coupling) correlation function with a model that does not include mode-coupling, whereas the RSD results fit the pre-reconstruction correlation function with a non-linear model that does include mode-coupling. Hence the expected values of α returned by both methods should be the same.

It is found that including such a prior has only a small effect on the recovered values and errors for $f\sigma_8$ and $b\sigma_8$, slightly decreasing the error range for both. The recovered best-fit values only change by a small amount compared to the statistical error on the measurements. This indicates that such a process introduces no bias into the results, which is not surprising, as the α prior comes from the comparison of the data itself before and after reconstruction, and systematic effects entering during the reconstruction process are expected to be very small. The reduction in the error range comes from the improvement in the Alcock-Paczynski measurement when the BAO position is known, and not from double counting as the adopted procedure carefully only includes the extra information recovered post-reconstruction.

4.3.2 Effects of $\sigma_{8,nl}$ Prior

The CLPT model's dependency on $\sigma_{8,nl}$ in the non-linear regime is weak enough that the MGS data provides no constraints on this except through the first order measurements of $b\sigma_8$ and $f\sigma_8$. The remaining non-linear contribution is largely unconstrained. A Planck+WP+highL (Planck Collaboration et al., 2014b) prior on $\sigma_{8,nl}$ is therefore considered, which takes the form of a Gaussian with mean $\sigma_{8,nl}(z_{eff}) = 0.766$ and variance 0.012. This stops the second order corrections to the model from straying into unphysical regions of parameter space, where the model itself is not expected to be accurate. For the baseline RSD fits, this prior is adopted, and is considered not to be introducing any additional information to the resultant measurements; rather it is simply forcing the fitting procedure to only consider physical solutions for the CLPT model.

When this prior is included there is a small change in the recovered mean values of

$f\sigma_8$ and $b\sigma_8$. For the average of the mocks the value of $f\sigma_8$ decreases slightly from 0.49 to 0.45. This shift actually brings the value of $f\sigma_8$ closer to that expected based on the cosmology used to generate the mocks and is well within the expected statistical deviation of the measurement. Additionally, adding in the $\sigma_{8,nl}$ prior increases the value of $b\sigma_8$ and tightens the constraints, bringing them closer to the expected value. This is because the prior allows constraints to be placed on the second order contribution to the galaxy bias, which, in the CLPT model, enters as additional small scale clustering proportional to $\langle F'' \rangle^2$. When this contribution is completely unconstrained, large values force the linear galaxy bias to be lower than it should be to fit the smallest scales. Due to the strong degeneracy between $b\sigma_8$ and $f\sigma_8$ it is actually this stronger constraint on $b\sigma_8$ that has a knock-on effect of reducing the value of $f\sigma_8$ obtained.

4.3.3 Testing Bin Width and Fitting Range

Several robustness tests, using the α and Planck prior measurement, are performed. The effects of changing both the bin width of the measurements and the fitting range are both considered. When the fitting range is reduced to $35\,h^{-1}\,\mathrm{Mpc} \leq s \leq 140\,h^{-1}\,\mathrm{Mpc}$ there is a slight increase in $f\sigma_8$, and a corresponding decrease in $b\sigma_8$, though these shifts are well within the statistical uncertainty. The reason for this shift stems from the higher order Lagrangian bias contributions: when the small scale data is removed, the constraints on $\langle F'' \rangle$ become much weaker and it is harder to decouple from $\langle F' \rangle$. The errors on $f\sigma_8$ and $b\sigma_8$ increase when we reduce the fitting range, consistent with the loss of information, particularly at small scales.

The results in Table 4.1 and Figure 4.5 also show that the choice of bin width has negligible effect on the results obtained. Cases 5 and 6 therein show fits using the fiducial fitting range and priors but using a correlation function and covariance matrix that has been binned using $\Delta s = 5\,h^{-1}\,\mathrm{Mpc}$ and $\Delta s = 10\,h^{-1}\,\mathrm{Mpc}$ respectively. The results are fully consistent with each other and the fiducial bin width case, with only small, statistically driven deviations in the mean and 1σ marginalised values of $f\sigma_8$ and $b\sigma_8$.

4.3.4 Effects of Fixing α and ϵ

Another case considered is for models where the values of α and ϵ are not varied, as in several previous studies (Blake et al., 2011a; Beutler et al., 2012; Samushia et al., 2012). This carries the implicit assumption that the fiducial cosmology is the true cosmology. Figure 4.7 shows the expected deviation of these parameters at the effective redshift of the MGS, assuming ΛCDM, based on the cosmological results from Planck (Planck Collaboration et al., 2014b). This uses the Planck ΛCDM base-planck-lowl-lowLike-highL

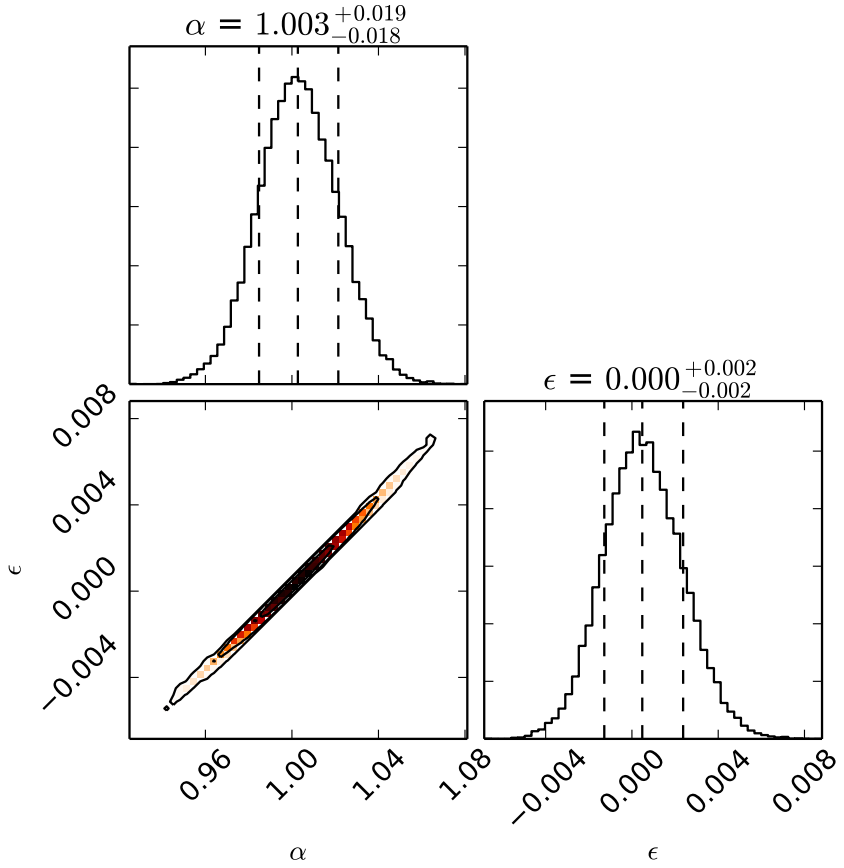


Figure 4.7: The 2D and 1D marginalised constraints on α and ϵ at $z = 0.15$ based on Planck ΛCDM cosmological constraints. Ellipses show the 1, 2 and 3σ regions, whilst dashed lines show the mean and 1σ errors of the marginalised distributions.

chains which is the basis for the fiducial cosmology adopted throughout this whole analysis.

It can be seen that ϵ, which is related to the AP parameter F as in Eq. 4.8, is very well defined at the effective redshift of the sample, with only a 1% deviation from $\epsilon = 0.0$ allowed to within 5σ. Even relatively large deviations from the fiducial cosmology manifest as only small changes in ϵ away from zero. As a majority of the information on ϵ comes from the quadrupole, which is also where most of the information on $f\sigma_8$ is obtained, it is concluded that the actual AP signal one would expect to measure as part of the RSD fit is also small.

However, from Figure 4.7, fixing alpha to the fiducial value is not supported by the Planck data, where even large deviations from $\alpha = 1.0$ can be found to within 5σ. It is mainly the monopole of the correlation function that constrains α, but the large degeneracies between α and $b\sigma_8$, and $b\sigma_8$ and $f\sigma_8$ mean that fixing this value could have

a knock-on effect on the $f\sigma_8$ constraints. As such it is hypothesised that though the expected degeneracy between the AP and RSD signals is small, not allowing α to vary could bias any constraints on $b\sigma_8$ and $f\sigma_8$.

Finally, it also important to note that Figure 4.7 is only true under the assumption of a ΛCDM cosmology. Allowing for $w_0 \neq 1.0$, a time-dependent equation of state for dark energy, or other non-standard cosmological models could allow for a much greater variation in α and ϵ from their fiducial values. As these phenomena are only emergent at late times they would be largely unconstrained by Planck, rendering any apparent Planck priors on α and ϵ moot.

To test this additional fits to the average of the mocks are made: first fixing $\epsilon = 0.0$ and allowing alpha to vary, then fixing ϵ and α. α is fixed to two different values: $\alpha = 1.00$, which is what is expected for the mean of the mocks, and $\alpha = 1.04$ which is the value recovered from the BAO-only fits to the MGS data.

Looking at Table 4.1 and Figure 4.5, the recovered values of $f\sigma_8$ and $b\sigma_8$ when fixing ϵ do shift slightly, but are still in very good agreement with the expected values for the mocks. This indicates that this is not introducing any bias into the results. The uncertainty on $f\sigma_8$ is also reduced substantially, with the lower bound especially reduced by a factor of 2. This is because confining the model to only those regions of parameter space that are in agreement with the Planck-ΛCDM predictions greatly reduces the degeneracy between $f\sigma_8$ and ϵ.

It should be noted however that this result would also be recovered if one were to take the case where α and ϵ are varied and simply combined with Planck data at a later stage, as the constraints from Planck are tight enough to effectively fix ϵ. Hence the benefit in allowing ϵ to vary at this stage, during the RSD fitting, is that the subsequent $f\sigma_8$ results are more general and can be combined with any additional models, not just those that agree with the Planck-ΛCDM constraints.

When fixing α to different values there is a small change in the recovered best fit values of $b\sigma_8$ and $f\sigma_8$, whilst the uncertainties therein remain unchanged. However this is not much beyond that seen when fixing ϵ to the value expected from the mocks. To reiterate, however, fixing α is not supported by the Planck-ΛCDM predictions and so this should be allowed to vary.

4.3.5 Using a Linear Model

The last case investigated is where a simple linear model, as opposed to the CLPT model, is used, as per Hamilton (1992). The results when using this model are shown in Table 4.1 and Figure 4.5. Here the reconstruction-motivated prior on α is still kept, and $f\sigma_8, b\sigma_8, \alpha, \epsilon$ and IC are the parameters allowed to vary. In this case the error budget for

both $f\sigma_8$ and $b\sigma_8$ is significantly reduced in comparison to the fiducial fit, and to a greater extent than when we use the perturbation theory model is retained but α and ϵ are fixed. A simple linear model neglects the contributions from higher order bias corrections which for the MGS are non-negligible and have been shown to affect the estimation of $b\sigma_8$ and, by way of the strong degeneracy therein, $f\sigma_8$. However, there is no significant bias in the recovered best-fit values themselves when using a linear model and any differences between the observed RSD signal and the prediction from linear theory are largely hidden by noise.

4.4 Growth Rate Measurements at $z = 0.15$ Using the MGS Data

This section presents the constraints on $f\sigma_8$ and $b\sigma_8$ from fitting the RSD model to the MGS data using the method tested in Sections 4.2 and 4.3. Therein, it has been shown that the fitting method is independent of the choice of priors, fitting range and bin size, but in the interest of completeness the same range of fits is performed on the MGS data itself. For equivalent fits to both data and mocks the covariance matrix is the same, so any differences stem from noise in the data or, of course, differences between the fiducial cosmology and the true cosmology. The marginalised mean values and 1σ constraints on $f\sigma_8$ and $b\sigma_8$ for all of the fits are given in Table 4.2 with the minimum χ^2 values, and plotted in the corresponding Fig. 4.8.

The fiducial fitting case including both α and $\sigma_{8,nl}$ priors is shown in Fig. 4.9, where the 2-D redshift space correlation function of the MGS data are plotted along with the maximum likelihood model. In Fig. 4.10, the recovered best-fit $b\sigma_8$-$f\sigma_8$ contour for the fiducial fitting case is plotted, alongside the marginalised 1D histograms for these parameters. This latter plot demonstrates the strong degeneracy between $f\sigma_8$ and $b\sigma_8$ that drives the small variations seen in the mean values when fitting to both the data and the average of the mocks.

For all the fits to the data it is worth noting that the model does seem to fit a slightly lower value for $b\sigma_8$ than would be expected based on the HOD fits to the MGS data. Looking back to Fig. 3.14 helps explain why. The amplitude of the monopole on the scales $25\,h^{-1}\,\mathrm{Mpc} \leq s \leq 60\,h^{-1}\,\mathrm{Mpc}$, where most of the information on the linear bias comes from, seems to be slightly lower for the data than for the HOD fit applied to mocks, though when scales above and below this range are included the mock amplitude is well matched. The fiducial fitting method is not including scales below $s = 25\,h^{-1}\,\mathrm{Mpc}$, where the mocks and data are in better agreement, and so it is not surprising the data prefers slightly smaller values of $b\sigma_8$.

Table 4.2: The mean values and one-sigma errors on $f\sigma_8$ and $b\sigma_8$ from fitting to the data monopole and quadrupole when different priors are applied and certain parameter combinations are fixed. For ΛCDM and GR the expectation is $f\sigma_8 = 0.466$ and, from the HOD fits to the MGS data, $1.15 \leq b\sigma_8 \leq 1.22$.

	Data:			
No.	Case	$f\sigma_8$	$b\sigma_8$	χ^2/dof
1	Full fit	$0.63^{+0.24}_{-0.27}$	$1.00^{+0.21}_{-0.19}$	26.0/26
2	prior on α	$0.64^{+0.23}_{-0.22}$	$0.98^{+0.16}_{-0.20}$	26.2/26
3	prior on $\sigma_{8,nl}$	$0.53^{+0.19}_{-0.19}$	$1.17^{+0.14}_{-0.18}$	28.6/26
4	$35 \leq s \leq 140\,h^{-1}\,\mathrm{Mpc}$	$0.56^{+0.25}_{-0.24}$	$1.08^{+0.14}_{-0.22}$	25.8/20
5	$\Delta s = 5\,h^{-1}\,\mathrm{Mpc}$	$0.52^{+0.19}_{-0.19}$	$1.16^{+0.13}_{-0.16}$	40.1/46
6	$\Delta s = 10\,h^{-1}\,\mathrm{Mpc}$	$0.49^{+0.17}_{-0.22}$	$1.19^{+0.12}_{-0.15}$	18.8/20
7	$\epsilon = 0.00$	$0.49^{+0.15}_{-0.14}$	$1.20^{+0.15}_{-0.15}$	31.0/27
8	$\alpha = 1.00$, $\epsilon = 0.00$	$0.44^{+0.16}_{-0.12}$	$1.12^{+0.09}_{-0.14}$	30.3/28
9	$\alpha = 1.04$, $\epsilon = 0.00$	$0.49^{+0.16}_{-0.13}$	$1.17^{+0.10}_{-0.12}$	31.0/28
10	Linear Fit	$0.47^{+0.13}_{-0.13}$	$1.15^{+0.08}_{-0.08}$	31.1/29

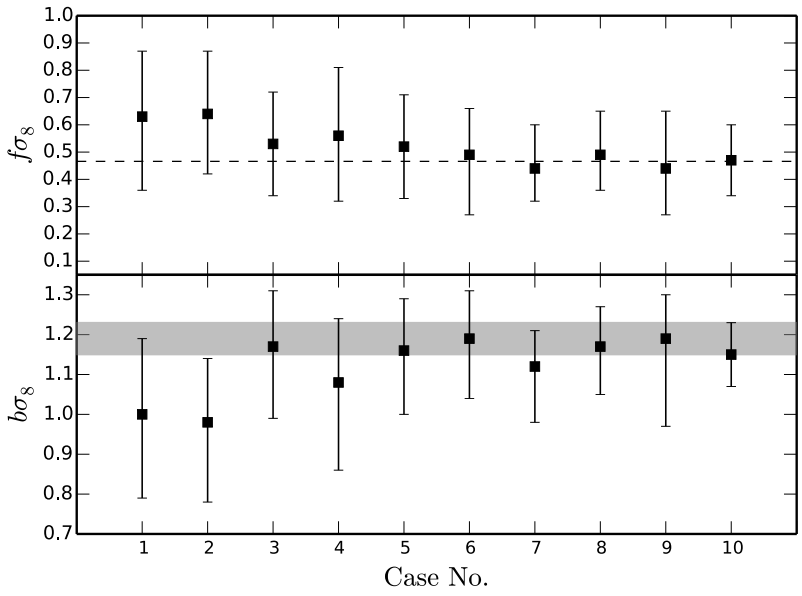


Figure 4.8: The marginalised $f\sigma_8$ and $b\sigma_8$ values and one-sigma errors from fitting to the data for the 10 cases listed in Table 4.2. As for Fig. 4.5, the dashed line indicates the expected growth rate assuming the fiducial ΛCDM cosmology. The shaded band indicates the expected linear galaxy bias as measured from the HOD fits to the MGS sample. A band rather than a line is used to account for the fact that the calculated value depends slightly on the range of scales used.

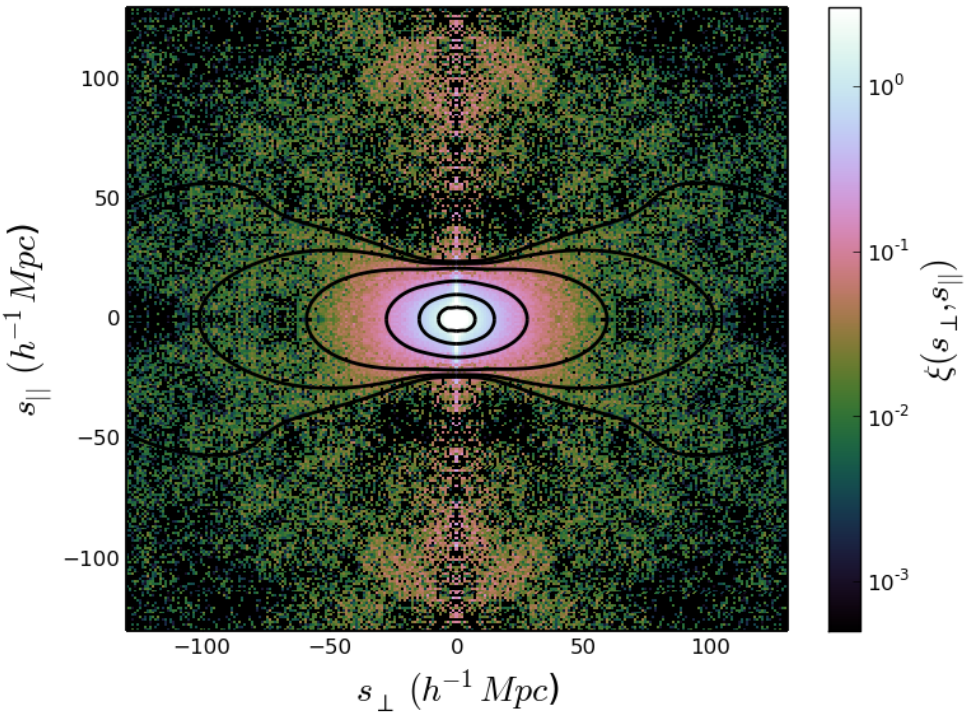


Figure 4.9: The 2D redshift space correlation function of the MGS along and perpendicular to the line of sight in bins of $\Delta s = 1\,h^{-1}\,\mathrm{Mpc}$. The solid black contours show the best-fit CLPT model at $\xi = \{0.001, 0.01, 0.04, 0.3, 2.0, 15.0\}$ using the fiducial fitting procedure.

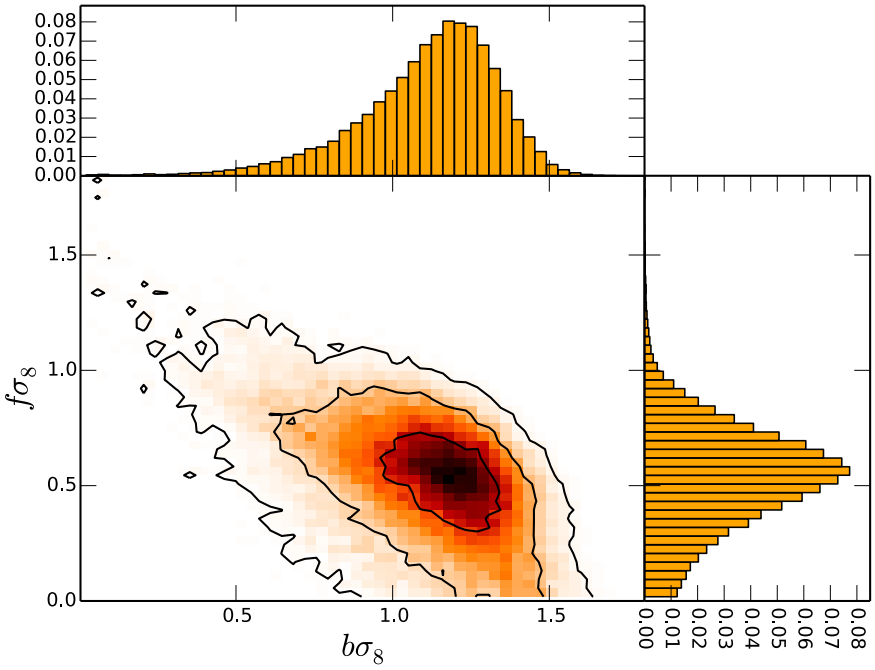


Figure 4.10: The 1, 2 and 3σ $b\sigma_8$ and $f\sigma_8$ likelihood contours and respective 1D marginalised likelihoods for the MGS using fits to the correlation function monopole and quadrupole in the range $25\,h^{-1}\,\mathrm{Mpc} \leq s \leq 160\,h^{-1}\,\mathrm{Mpc}$ with bins of width $\Delta s = 8\,h^{-1}\,\mathrm{Mpc}$ and priors on α and $\sigma_{8,nl}$.

4.4.1 Effects of α Prior

As for the results fitting the average of the mocks, adding a prior on α introduces no noticeable bias in the best fit $f\sigma_8$ and $b\sigma_8$ values and only a slight reduction in the errors. When fitting to the data, the best fit χ^2 increases slightly from 26.0 to 26.2 for 26 degrees of freedom (34 bins and 8 free parameters) when a prior on α is introduced. Such an increase is to be expected as the prior forces the best-fit model away from the overall maximum likelihood model. However, the difference is very small, indicating no strong preference for models outside the prior range.

4.4.2 Effects of $\sigma_{8,nl}$ Prior

When a Planck prior on $\sigma_{8,nl}$ is added, there is a larger difference in the $f\sigma_8$ and $b\sigma_8$ constraints than for the mocks, though the value of $f\sigma_8$ does not shift by more than would be expected based on the statistical errors. As has been demonstrated previously in this chapter, this prior is not believed to be adding any bias to the results from the tests on the mocks and as such any change in the recovered constraints is purely statistically driven. Before adding in the $\sigma_{8,nl}$ prior the measured values of $b\sigma_8$ are lower than one would expect, but this value increases by $\sim 1\sigma$ when this prior is included. It is this change in the mean recovered value of $b\sigma_8$ which causes the slight change in $f\sigma_8$. The reason for the underestimation of $b\sigma_8$ is as mentioned previously; without this prior helping to constrain $\sigma_{8,nl}$, $\langle F'' \rangle$ is overestimated and $b\sigma_8$ underestimated. For this prior the best-fit $\chi^2 = 28.6$, which is again a slight increase compared to the fits with only the α prior, however for all three cases with different priors the recovered χ^2 values are very reasonable.

4.4.3 Effects of Different Bin Widths and Fitting Ranges

When the fitting range or the bin size is changed, the results remain similar to those of the fiducial case, and as with the average of the mocks there is no indication that the fiducial fitting choices are creating biased results. As for the average of the mocks, removing the smallest scales from the fits reduces the recovered $b\sigma_8$ value and increases the error, but the mean $f\sigma_8$ remains almost unchanged. For all of the tests of bin width and fitting range, the χ^2 values are in agreement with the fiducial case, which indicates that all of the fits are good. The largest χ^2/dof belongs to the case where the fitting range is modified. Here, $\chi^2 = 25.8$ for 20 degrees of freedom. However, this value is still very good and a worse χ^2 could be expected approximately 17% of the time.

4.4.4 Effects of Fixing α and ϵ or Using a Linear Model

The final set of fits performed, fixing α and ϵ and using a simpler linear model, corroborate the results when fitting to the average of the mocks. Making use of the reasonable assumption that $\epsilon = 0.0$ tightens the constraints on $b\sigma_8$ and $f\sigma_8$ without adding any notable change in the best fit results. The upper and lower bounds on $f\sigma_8$ reduce from 0.19 and 0.19 to 0.15 to 0.14 respectively. Fixing α to different values does change the best fit results slightly too, as was seen in the fits to the mean of the mocks, whilst keeping the errors almost unchanged compared to the fixed ϵ case. This is not a substantial change, though as there are not strong Planck constraints on α, (unlike for ϵ), it is concluded that fixing α could lead to biased results.

Overall, the χ^2 values found when fixing α and ϵ or using a linear model are similar in comparison to using the CLPT model and allowing α and ϵ to vary. The data is not powerful enough to discriminate between these different models, however from Wang et al. (2014) it cannot be expected that a linear model is able to fully reproduce the RSD signal on the smallest scales that are fit against. At these scales non-linear effects start to dominate, and when fitting the RSD signal on these scales the CLPT model is a more reliable choice.

4.4.5 Comparison of Different MGS Results

A range of RSD fits have been made to the MGS data, assuming different values for α and ϵ. This section provides an overview of those that are quoted as the'final' result, those that should be used for further cosmological studies and those that should not.

By fitting the full-shape of the correlation function monopole and quadrupole, and varying α and ϵ, the best-fit values are $f\sigma_8 = 0.53^{+0.19}_{-0.19}$ and $b\sigma_8 = 1.17^{+0.14}_{-0.18}$. These values make no assumption on the underlying, late-time, cosmology and so are recommended for future cosmological constraints. These will be used in Section 4.5.3 to constrain the growth index, γ, and compare this to the prediction from General Relativity. As the 1-D $f\sigma_8$ and 3-D $f\sigma_8, \alpha$ and ϵ likelihoods cannot be well approximated by a Gaussian, the likelihoods themselves are used to achieve this, rather than just the quoted numbers. For future analyses making use of the MGS RSD results the prepared MCMC samples for this fit have been made publicly available.

If a ΛCDM cosmology is assumed, the RSD constraints can be improved by fixing $\epsilon = 0.0$ yet still allowing α to vary. Here $f\sigma_8 = 0.49^{+0.15}_{-0.14}$ and $b\sigma_8 = 1.20^{+0.15}_{-0.15}$. This is well motivated by the Planck data, where, unless one adopts a late time dark energy model quite different from those commonly assumed, there is no expected, detectable deviation from $\epsilon = 0.0$. As such this measurement is presented as the quoted, fiducial result and should be used for comparison with other $f\sigma_8$ results under the ΛCDM framework.

However, this result should not be combined with Planck data as that would result in effectively double counting the Planck constraints. Rather, from Figure 4.7, combining the publicly available chains with Planck data will effectively fix ϵ and recover the fiducial results. From the same figure though it is not recommended that the results where α is not allowed to vary are used. In fact, as α dilates the whole correlation function, not just the BAO peak, and captures the late-time cosmological dependence of the shape of the correlation even on small scales, it is recommended that α be allowed to vary for any measurements of the growth of structure.

4.5 Cosmological Interpretation

4.5.1 BAO Distance Ladder

The MGS BAO measurement from Section 4.1.3 provides a new rung for the BAO distance ladder. Fig. 4.11 shows current BAO-scale measurements compared with the MGS result, displayed using a red diamond. The measurements in Fig. 4.11 are divided by the prediction for the best-fit flat ΛCDM model, as determined from the Planck satellite (Planck Collaboration et al. 2014a,b) observations of the CMB. The grey contour represents the 1σ allowed region, determined by sampling the same Planck likelihood chains used for the RSD measurement priors and associated plots. The points displayed using black circles form a set of independent measurements; these include the 6dFGS measurement made by Beutler et al. (2011) and BOSS measurements made by Tojeiro et al. (2014) and Anderson et al. (2014b). This data is combined with the MGS measurement to obtain cosmological constraints in the following section. The white squares represent measurements made using SDSS DR7 data (Percival et al. 2010 and Xu et al. 2012) and the grey squares are measurements made by Kazin et al. (2014) using WiggleZ data. These data overlap significantly in volume with the BOSS data and as such are not used in this chapter to obtain cosmological constraints. The BAO distance measurements are broadly consistent with each other and the Planck best-fit ΛCDM prediction.

4.5.2 Cosmological Constraints with BAO

In total, four independent galaxy BAO measurements have been identified that can be combined to obtain cosmological constraints. These include three spherically-averaged measurements; the MGS measurement at $z = 0.15$, Beutler et al. (2011) at $z = 0.11$, Tojeiro et al. (2014) at $z = 0.32$, and the anisotropic measurement of Anderson et al. (2014b) at $z = 0.57$. From here on, 'BOSS' denotes that the (Tojeiro et al., 2014) and (Anderson et al., 2014b) anisotropic BAO measurements are used. '6dF' denotes that the 6dFGS BAO measurement of Beutler et al. (2011) is used. These data are com-

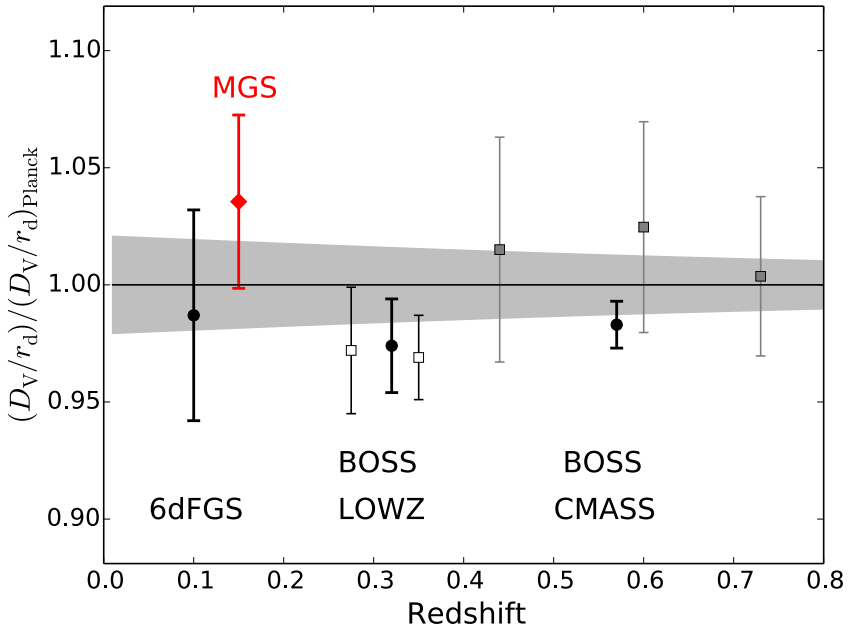


Figure 4.11: The BAO distance ladder, expressed as D_V/r_d, including the MGS measurement and relative to the Planck prediction given their best-fit flat ΛCDM model. The grey region represents the 1σ uncertainty given Planck data and assuming a flat ΛCDM model. The MGS measurement, using the SDSS DR7 main galaxy sample, is displayed with a red diamond. Measurements made using 6dFGS data (Beutler et al., 2011) and BOSS data (Anderson et al. 2014b; Tojeiro et al. 2014) are denoted with black circles. These measurements are nearly independent with those presented in this chapter, allowing them to be combined to obtain cosmological constraints. The white squares display measurements using SDSS DR7 data (Percival et al. 2010; Xu et al. 2012) and the grey squares display measurements made using WiggleZ data (Kazin et al., 2014).

bined with the CMB results released by Planck Collaboration et al. (2014b) that are based on the combination of data from the Planck Satellite, Wilkinson Microwave Anisotropy Probe (WMAP) satellite (Bennett et al., 2003; Spergel et al., 2003) polarization measurements (Bennett et al., 2013), and high-ℓ power spectra data from ACT (Das et al., 2014) and SPT (Story et al., 2013) and denoted 'Planck+WP+highL' in Planck Collaboration et al. (2014b). Here and previously this combination of CMB data has simply been referred to as 'Planck'. Likelihoods for cosmological parameters are determined for the BAO+Planck data set using the COSMOMC software package (Lewis et al., 2002; Lewis , 2013). A study by Aubourg et al. (2014) explores the constraints that are achieved when also including BOSS Lyman α forest BAO measurements (Font-Ribera et al., 2014; Delubac et al., 2014) and when considering many extensions to the basic ΛCDM model. Here, only simple extensions of the ΛCDM cosmological model are considered.

When the equation of state of dark energy is allowed to vary, the MGS BAO measurement provides significant improvement in the precision of Ω_m, H_0, and w_0 compared to the case of Planck+BOSS+6dF measurements. Adding the MGS measurement to either the Planck+BOSS or Planck+BOSS+6dF data sets results in a 15 per cent improvement in the precision of the H_0 and w_0 measurements. This is illustrated in Fig. 4.12, where the 1 and 2σ allowed regions for w_0 and H_0 are displayed for Planck+BOSS (red) and Planck+BOSS+ MGS (green).

Using the full data set, a value of $w_0 = -1.010 \pm 0.081$ is found and the best-fit cosmological parameters differ from the the Planck ΛCDM best-fit by less than 0.4σ (where σ is the uncertainty on the Planck best-fit measurements). The new MGS BAO measurement thus affords significant improvement in measurements of the properties of dark energy and, in combination with other BAO data, is in excellent agreement with a flat ΛCDM model.

In all of the cases compared, H_0 decreases when the MGS measurement is included. For example, in the case of varying curvature and dark energy equation of state, $H_0 = 67.5 \pm 2.2\,\mathrm{kms}^{-1}\mathrm{Mpc}^{-1}$ for the combination of Planck+BOSS+MGS data, while excluding the MGS measurement yields $H_0 = 69.3 \pm 2.8\,\mathrm{kms}^{-1}\mathrm{Mpc}^{-1}$. The MGS BAO measurement therefore increases the tension between Planck+BAO measurements of H_0, and those obtained using direct detection, e.g., the measurements by Efstathiou (2014) of $H_0 = 72.5 \pm 2.5\,\mathrm{kms}^{-1}\mathrm{Mpc}^{-1}$, Riess et al. (2011) of $H_0 = 73.8 \pm 2.4\,\mathrm{kms}^{-1}\mathrm{Mpc}^{-1}$, and Freedman et al. (2012) of $H_0 = 74.3 \pm 2.1\,\mathrm{kms}^{-1}\mathrm{Mpc}^{-1}$. The constraint on H_0 obtained using BAO measurements is explored in much greater detail in the study by Aubourg et al. (2014).

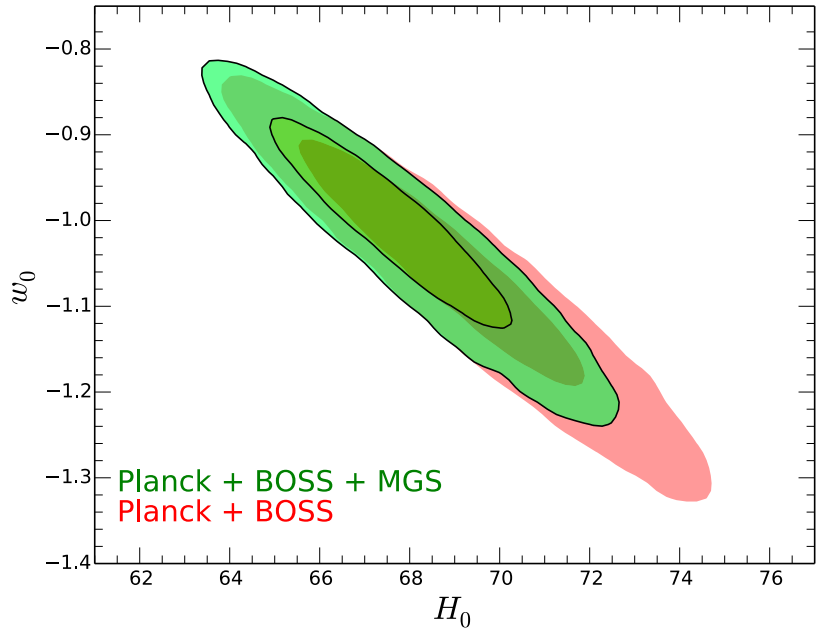


Figure 4.12: The 1 and 2σ confidence levels for the dark energy equation of state, w_0, and the value of the Hubble constant, H_0, constraints combining BAO distance measurements with Planck data. The results when including Planck and BOSS data are shown in red and the result including the new measurement made using SDSS DR7 MGS data in green. The inclusion of the MGS measurement decreases the area enclosed by the 1σ contour by 20 per cent. This figure is taken from Ross et al. (2015)

4.5.3 Cosmological Interpretation of RSD measurements and Comparison to Previous Studies

This section compares the MGS measurement of $f\sigma_8$ with those from a range of different galaxy surveys and performs a simple consistency test against the prediction of the growth rate from General Relativity (GR). This uses the common γ parametrisation of the growth rate, where $f(z)$ is approximated as

$$f(z) = \Omega_m(z)^{\gamma}. \tag{4.12}$$

For GR, $\gamma \approx 0.55$ (Linder & Cahn, 2007).

Measurements of $f\sigma_8$ have been made up to $z = 0.8$ using data from the 2-degree Field Galaxy Redshift (2dFGRS; Percival et al. 2004), 6-degree Field Galaxy (6dFGS; Beutler et al. 2012), SDSS-II Luminous Red Galaxy (Samushia et al., 2012; Oka et al., 2014), BOSS (Beutler et al., 2013; Chuang et al., 2013; Samushia et al., 2014; Sánchez et al., 2014), VVDS (Guzzo et al., 2008) and WiggleZ (Blake et al., 2011a,b) surveys among others. Although these measurements were all made using different models of varying complexity and different fitting methods to either the correlation function or power spectrum, they can be roughly grouped into two distinct categories: those that were made assuming a fixed fiducial cosmological model and those that fit the full shape of the galaxy clustering statistics. The latter simultaneously measures both the RSD and BAO signals and as such includes the degeneracy between $f\sigma_8$, $b\sigma_8$ and α highlighted in Section 4.3.4

These two compilations of measurements are plotted separately in Fig. 4.13. The $z = 0.57$ BOSS and four WiggleZ measurements were calculated with and without the inclusion of the AP effect. They too find a large difference in the constraints when incorporating this degeneracy into their measurements. Also plotted alongside these measurements are the Planck-ΛCDM predictions for $f\sigma_8$ assuming different values for the γ parameter. The majority of the measurements, including the new MGS measurements, are in good agreement with the GR prediction.

As a more quantitative consistency test of GR the likelihoods recovered from the full-fit MCMC analysis of the MGS are used to put constraints on γ itself. The MGS RSD result is combined with the publicly available Planck likelihood chains, subsampled to enforce a prior on Ω_m. The Planck chain is importance-sampled by randomly choosing a value $0 \leq \gamma \leq 1.5$ for each point in the chain and evaluating the likelihood for that parameter combination. One caveat, however, is that the Planck value of σ_8 has to be corrected to account for the fact that this also depends on γ. Each point in the Planck chain gives a value of $\Omega_{m,0}$ and $\sigma_{8,0}$, where the latter is derived from the CMB power spectrum amplitude assuming GR. The correct value of $f\sigma_8$ is then evaluated by scaling back σ_8 to a suitably high redshift (for simplicity the redshift of recombination $z*$ is used,

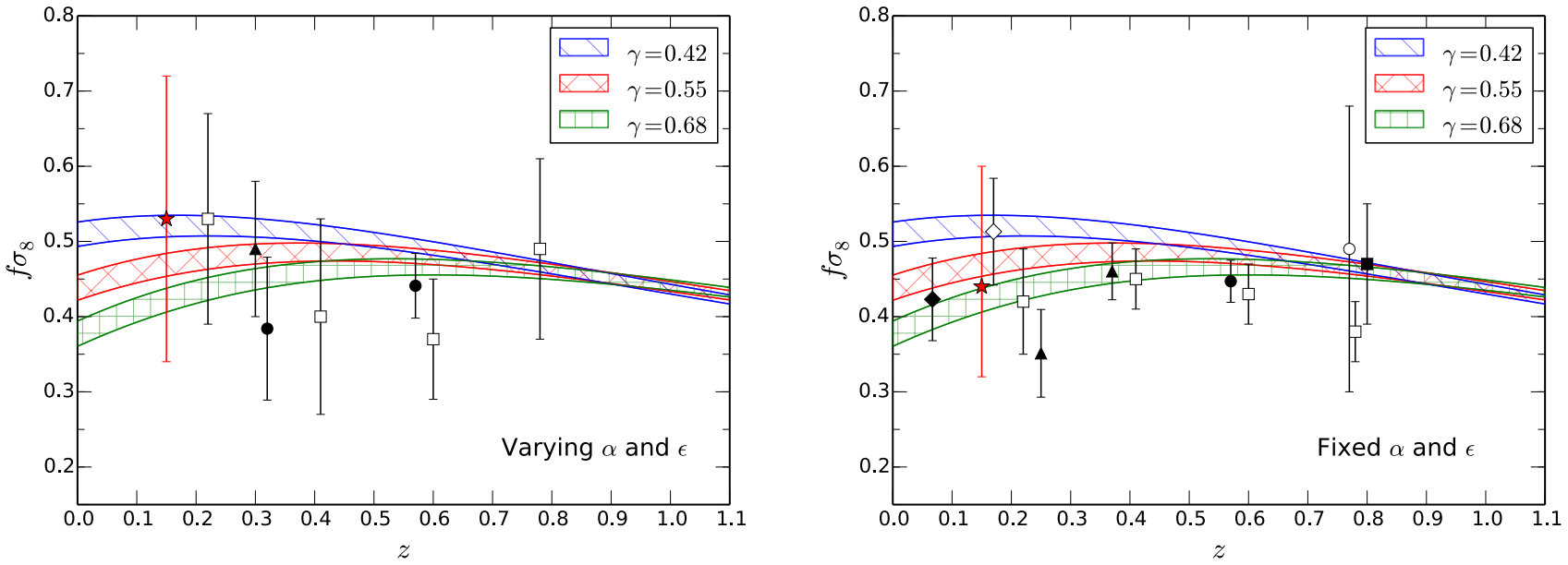


Figure 4.13: A comparison of measurements of the growth rate using the two-point clustering statistics from a variety of galaxy surveys below $z = 0.8$. The results are split into two groups: those that perform a full shape fit, varying α and ϵ; and those that just fit the growth rate for a fixed cosmology, neglecting the degeneracy between α, $b\sigma_8$ and $f\sigma_8$. The MGS RSD measurement is shown as a filled red star, with other data points representing the 6dFGS (filled diamond; Beutler et al. 2012), 2dFGRS (empty diamond; Percival et al. 2004), SDSS-II LRG (filled triangle; Samushia et al. 2012 (no AP), Oka et al. 2014 (AP)), BOSS (filled circle; Chuang et al. 2013 (z=0.32), Samushia et al. 2014 (z=0.57)), WiggleZ (open square; Blake et al. 2011a,b), VVDS (open circle; Guzzo et al. 2008) and VIPERS (filled square; de la Torre et al. 2013) surveys. Also included are Planck predictions for the growth rate for values of $\gamma = 0.42$, 0.55 and 0.68 as hatcheds bands (top, middle and bottom respectively).

which is also given at each point in the Planck chain) and then scaling both σ_8 and Ω_m to the effective redshift of the MGS using the correct value of γ. i.e., for scale factor $a = 1/(1+z)$,

$$f(a)\sigma_8(a) = \Omega_m(a)^\gamma \sigma_{8,0} \frac{D_{gr}(a*)}{D_{gr,0}} \frac{D_\gamma(a)}{D_\gamma(a*)} \tag{4.13}$$

where,

$$\Omega_m(a) = \frac{\Omega_{m,0}}{a^3 E(a)^2} \tag{4.14}$$

$$D_{gr}(a) = \frac{H(a)}{H_0} \int_0^a \frac{da'}{a'^3 H(a')^3} \tag{4.15}$$

$$\frac{D_\gamma(a)}{D_\gamma(a*)} = \exp\left[\int_{a*}^{a} \Omega_m(a')^\gamma d\ln a'\right] \tag{4.16}$$

and

$$H(a) = H_0 E(a) = H_0 \sqrt{\frac{\Omega_{m,0}}{a^3} + \frac{(1-\Omega_{m,0}-\Omega_{\Lambda,0})}{a^2} + \Omega_{\Lambda,0}} \tag{4.17}$$

Even though the fiducial $f\sigma_8$ measurements use a prior to better constrain $\sigma_{8,nl}$ and stop the non-linear aspects of the CLPT model from straying into non-physical regions of the cosmological parameter space, all of the information on $f\sigma_8$, α and ϵ comes solely from the amplitude and BAO features of the monopole and quadrupole. As such the MGS RSD results can be combined with Planck data for this consistency test without the risk of double counting the Planck measurements.

The subsequent constraints on γ and Ω_m are shown in Fig. 4.14. Also shown are the joint constraints when including the measurements of $f\sigma_8$ from the BOSS-DR11 CMASS sample (Samushia et al., 2014). For this simple consistency check only the CMASS measurement is included as the method used to make this measurement is very similar to that used in this work. On top of this, the BOSS-DR11 LOWZ and WiggleZ measurements do overlap partially in terms of area and redshift distribution with both the MGS measurement and the CMASS measurement, so to properly include these would require an accurate computation of the cross correlation between these measurements which is beyond the scope of the work presented in this chapter. When combining the MGS result with the Planck prior $\gamma = 0.58^{+0.50}_{-0.30}$ is recovered, consistent with GR. With the addition of the CMASS measurement the constraints tighten to $\gamma = 0.67^{+0.18}_{-0.15}$, which is also consistent with GR to within 1σ. However it should be noted that in both cases there is a slight preference for higher values of γ than would be expected from GR.

This is taken one step further by including BAO information from the anisotropic RSD measurement and from the BOSS-DR11 CMASS results. The inclusion of anisotropic distance information helps to better constrain Ω_m and hence can reduce the uncertainty on γ constraints. The 3D $f\sigma_8, \alpha$ and ϵ likelihood from the fiducial MGS RSD fits are used as well as the equivalent constraints from the CMASS sample. The results of this

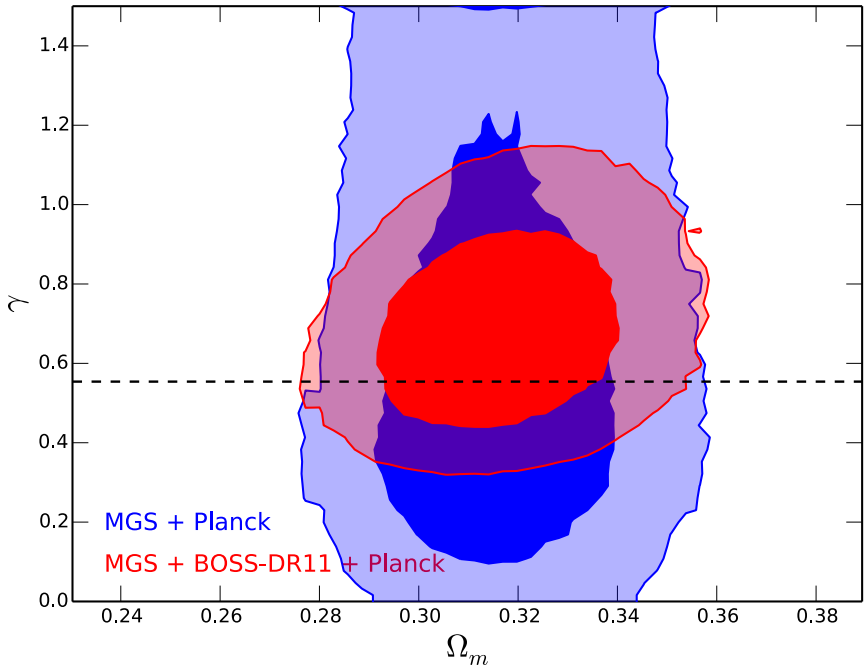


Figure 4.14: Constraints on γ and Ω_m from the combination of the marginalised MGS $f\sigma_8$ and Planck likelihoods. Contours correspond to the 1σ and 2σ confidence intervals of the recovered posterior distribution. The inclusion of the BOSS-DR11 CMASS measurement of the growth rate (Samushia et al., 2014) is also considered. In both cases there is good agreement with the prediction from GR (dotted line).

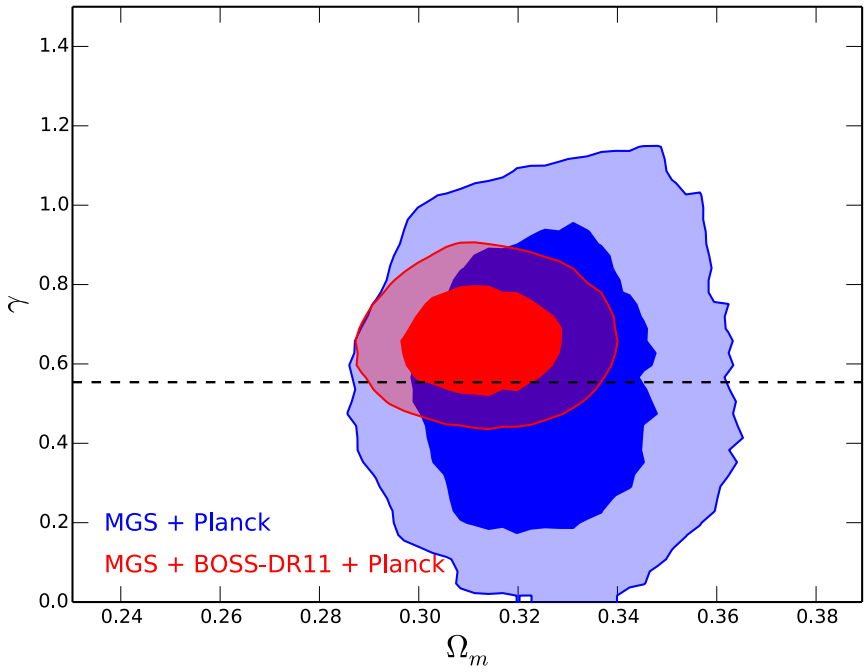


Figure 4.15: Constraints on γ and Ω_m from the combination of the 3-dimensional, marginalised MGS $f\sigma_8$, α and ϵ likelihood with the Planck likelihood. Contours correspond to the 1σ and 2σ confidence intervals of the recovered posterior distribution. In both cases there is good agreement with the prediction from GR (dotted line) and a reduction in the uncertainty on γ, compared to Fig. 4.14, when the anisotropic BAO information from the CMASS and MGS measurements is included.

are shown in Fig. 4.15, where a value of $\gamma = 0.64 \pm 0.09$ is found with the inclusion of the CMASS measurement, and $\gamma = 0.54^{+0.25}_{-0.24}$ is found without. Both of these measurements are consistent with GR to within 1σ. The addition of the MGS $f\sigma_8, \alpha$ and ϵ measurements improves the constraints on γ by $\sim 10\%$ compared to the constraints obtained using the CMASS measurement alone.

The growth index has also been measured by Beutler et al. (2013), Sánchez et al. (2014) and Samushia et al. (2014) from the combination of BOSS CMASS and Planck data. Additionally Sánchez et al. (2014) use BOSS LOWZ data to produce their constraints. In Fig. 4.16 the MGS+Planck constraint on γ is plotted alongside these other measurements. There is good consistency between all measurements, even though the methods used to measure the growth rate and anisotropic BAO information are very different. In all cases there is also a slight preference for higher values of γ, which corresponds to models where gravitational interactions are weaker.

There exists significant tension ($\sim 2.3\sigma$) between the Beutler et al. (2013) BOSS CMASS measurement of the growth index and the prediction from GR. An interesting

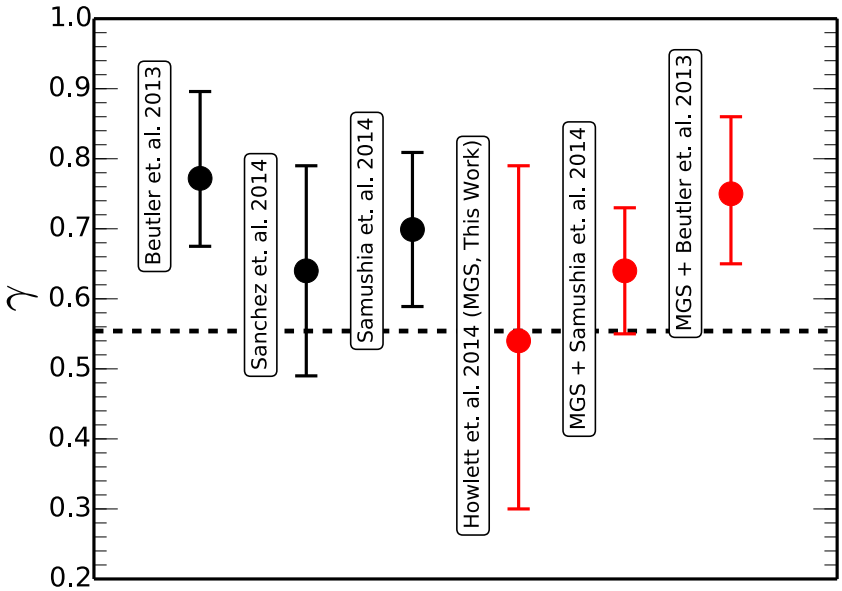


Figure 4.16: A comparison of γ constraints from several independent measurements of the growth rate using combinations of BOSS CMASS (and in the case of Sánchez et al. 2014, BOSS LOWZ) and Planck data. For consistency the MGS+Planck only measurement is plotted alongside. There is good agreement between all independent probes and a somewhat consistent favour for higher values of γ than would be predicted by GR (dashed line).

question to ask is whether the addition of the MGS measurement at low redshift helps to alleviate this tension and how this combination of measurements compares to the result presented previously when combining the MGS and Samushia et al. (2014) BOSS CMASS measurements. The results from these two combinations are also presented in Fig. 4.16, where one can see that the MGS measurement brings both combinations towards better agreement with the GR prediction, however there is still a 2σ tension between this prediction and the value of γ recovered when combining the MGS measurements with the Beutler et al. (2013) CMASS results.

4.6 Summary

This chapter has demonstrated the use of the MGS dataset and mocks, which were motivated and described in Chapter 3, as tools for a new set of low redshift cosmological constraints. The models used to measure the BAO and RSD features in the data have been described and tested on the mock catalogues before being applied to the MGS dataset itself.

For the BAO measurements a model similar to that of Anderson et al. (2014b) was used and a reconstruction algorithm applied to the data improves the precision of the BAO measurement by greater than a factor of two for both Fourier- and configuration-space measurements. These measurements were shown to be robust against choices in the fitting methodology. The BAO Fourier- and configuration-space measurements are consistent with each other and are combined to obtain the consensus measurement of $D_V(z_{\rm eff} = 0.15) = (664 \pm 25)(r_d/r_{d,\rm fid})$ Mpc. Combining this distance scale measurement with Planck CMB data and other BAO distance scale measurements improves the precision of cosmological constraints. For example, including the MGS BAO measurement in addition to BOSS and 6dFGS measurements improves the precision on the equation of state of dark energy by 15 per cent, to $w_0 = -1.010 \pm -0.081$.

The RSD measurements make use of the state-of-the-art CLPT model (Wang et al., 2014) to fit the monopole and quadrupole of the correlation function. This model was tested and verified on the MGS mock catalogues before being applied to the data. The resultant growth rate measurements, robust to changes in the fitting methodology, are $f\sigma_8 = 0.53^{+0.19}_{-0.19}$ when fitting to the full shape of the correlation function and $f\sigma_8 = 0.49^{+0.15}_{-0.14}$ when assuming no AP effect and fixing $\epsilon = 0$. This latter assumption is validated by the fact that it is expected one would detect $\epsilon = 0$ for any commonly assumed model of the expansion history based on the Planck-ΛCDM results.

Finally, as a consistency test of General Relativity, the growth index, γ, is fit. The results are $\gamma = 0.58^{+0.50}_{-0.30}$ when including Planck data and $\gamma = 0.67^{+0.18}_{-0.15}$ when also including BOSS-DR11 CMASS measurements of the growth rate. Including the additional

anisotropic BAO from the full fits to the shape of the correlation function our constraints tighten to $\gamma = 0.54^{+0.25}_{-0.24}$ and $\gamma = 0.64 \pm 0.09$ respectively, the latter of which is a $\approx 10\%$ improvements on the constraints from the CMASS and Planck measurements alone. All of these results are fully consistent with the predictions of General Relativity.

Chapter 5

Optimal Covariance Matrix Estimation for Next Generation Surveys

For random-phase, Gaussian distributed density perturbations, all the cosmological information is included in the 1- and 2-point correlation functions and the covariance matrix. Although gravitational evolution (and, if it exists, primordial non-Gaussianity) introduces small higher-order n-point functions, the majority of available information is still encapsulated in just the power spectrum and its associated covariance matrix. The former of these can be readily measured using large surveys of the universe, as demonstrated in Chapter 4, however the covariance matrix, which encodes the true underlying distribution from which the universe's power spectrum is drawn cannot be measured so easily. It must be modelled in some fashion. This can be done using large ensembles of fast, approximate simulations, as per Chapters 2 and 3. However, as LSS surveys continue to grow, running large numbers of simulations that have enough volume and are accurate enough becomes challenging.

This chapter presents a new optimal method of determining the covariance matrix using ensembles of simulations by combining analytic estimates of the covariance matrix with an ensemble of small volume simulations, hence bypassing the need to simulate regions of the universe that are large enough to encompass the survey volume, or to rely on complex theoretical models of non-linear galaxy clustering. The motivation behind this method will be further detailed in Section 5.1. Sections 5.2 and 5.3 will present a derivation of the FKP power spectrum estimator, briefly introduced in Chapter 1, before using the same techniques to derive an analytic expression for covariance matrix, which includes all possible terms and the effect of the survey window function.

As a test of the fidelity of this expression, Section 5.4 presents a derivation of the small

scale limit of the covariance from the new formula and compares this to existing expressions in the literature. Section 5.5 goes further and presents the expression in the absence of a survey window function, i.e., as would be measured from simulations. Section 5.6 then shows additional corrections that must be applied to correct for the 'supersample' covariance which is absent in simulations compared to a true survey due to the lack of modes larger than the simulation box.

The remainder of this chapter is devoted to combining this expression with simulations to generate the covariance matrix for a set of large simulations from a set of smaller simulations. This technique is first presented in Section 5.7. This section also shows that the new combined method including the supersample covariance correction recovers the small scale covariance matrix from the masked MGS galaxy mocks used in Chapter 3 and 4, using only cubic galaxy mocks 1/8 of the size.

This chapter is summarised in Section 5.8 which also gives examples of how the technique presented in this chapter can be applied to current and next generation surveys.

5.1 Motivation

Modelling the covariance matrix of two-point clustering is one of the most computationally demanding aspects of modern large scale structure analysis. Although this can be calculated analytically in the linear regime (Tegmark, 1997), the non-linear galaxy covariance matrix is a complex function of non-linear shot-noise, galaxy evolution and the unknown relationship between the galaxies and the underlying dark matter. In any real application this is further compounded by the effect of RSD. As such, a much more common solution is to use a set of detailed galaxy simulations, otherwise known as mock catalogues (mocks), to either fully estimate the covariance matrix or as the basis for an empirically motivated analytic fit (Xu et al., 2012).

In most recent analyses this estimation was performed using large numbers of simulations that cover the full survey volume, in both the angular and radial sense, and are accurate enough to reproduce the halo mass function down to the smallest halos in which the galaxies within the survey are found. Reaching the desired non-linear accuracy for future surveys such as the Large Sky Synoptic Telescope (LSST; Ivezic et al. 2008) and Euclid (Laureijs et al., 2011), may require using more complex computational methods than current surveys. It is reasonable to assume these methods will take much longer per simulation and hence the computational time required to generate an ensemble will be much greater for next generation surveys.

Furthermore, the particle mass within a simulation is given by

$$M_{\rm part} = \frac{\Omega_{m,0}\rho_c V}{N_{\rm part}}, \tag{5.1}$$

and hence for increasingly large volumes, the number of particles must increase in conjunction to recover the same halo mass. As an example, take the Euclid survey, which roughly covers the redshift range $1.0 < z < 2.0$ with a sky coverage of $15,000\mathrm{deg}^2$. This corresponds to a cosmological volume $\sim 5\times10^{10}\,h^{-3}\,\mathrm{Mpc}^3$. With the above formula, assuming 10 particles per halo, reaching a halo mass of $10^{11}M_\odot$ would require simulations containing approximately 7500^3 particles! Even if computing power improves rapidly over the next decade running a single simulation this size presents a daunting challenge. Running a large ensemble will be virtually impossible. One could imagine running separate simulations for each redshift slice in the survey, however this then increases the number of simulations required very quickly and does not truly fix this problem.

This is also not the only concern. Not only will future surveys require larger, more detailed simulations than ever before, but the number of simulations in an ensemble will also need to increase compared to current surveys. For current surveys, recent studies by Dodelson & Schneider (2013); Taylor et al. (2013) and Percival et al. (2014) have shown that $\mathcal{O}(1000)$ mocks are required to obtain an accurate numerical estimate of the covariance matrix with sub-dominant errors compared to the statistical errors themselves. However as the statistical errors in the surveys decrease, the number of simulations must increase to ensure the precision on the covariance matrix remains subdominant.

If the density perturbations present in the universe are drawn from a Gaussian distribution then the estimated power spectrum must be drawn from a chi-squared distribution and the estimated covariance matrix from its higher-dimensional counterpart, the Wishart distribution,

$$\mathcal{P}(\hat{\boldsymbol{C}}|\boldsymbol{C},n,p) = \left(\frac{n^{\frac{np}{2}}|\hat{\boldsymbol{C}}|^{\frac{n-p-1}{2}}}{2^{\frac{np}{2}}|\boldsymbol{C}|^{-\frac{n}{2}}\Gamma_p\left(\frac{n}{2}\right)}\right)e^{-\frac{1}{2}\mathrm{Tr}[\hat{\boldsymbol{C}}\boldsymbol{C}^{-1}]}, \tag{5.2}$$

where $\mathcal{P}$ gives the probability of measuring a $p\times p$ covariance matrix $\hat{\boldsymbol{C}}$ based on the true underlying covariance matrix $\boldsymbol{C}$. n is the number of degrees of freedom, which in the case of covariance matrices estimated from a set of mocks is $n = N_s - N_b - 1$, where N_s is the number of simulations and N_b is the number of measurement bins. Γ_p is the multivariate gamma function.

The covariance of the Wishart distribution and hence in the measured covariance matrix, is given by

$$\langle\Delta\hat{\boldsymbol{C}}_{i,j}\Delta\hat{\boldsymbol{C}}_{k,m}\rangle = n^{-1}(\boldsymbol{C}_{i,k}\boldsymbol{C}_{j,m} + \boldsymbol{C}_{i,m}\boldsymbol{C}_{j,k}). \tag{5.3}$$

In the simplified case of a Gaussian random field where the covariance matrix is diagonal, this reduces to

$$\Delta\hat{\boldsymbol{C}}_{i,i} = \sqrt{\frac{2}{n}}\boldsymbol{C}_{i,i}. \tag{5.4}$$

Hence the error on the covariance matrix scales as one over the square root of the number of degrees of freedom. This scaling has been tested and verified even for non-linear

simulations by Takahashi et al. (2009) and Takahashi et al. (2011). For a number of mocks much larger than the number of measurement bins, this then means that the precision of the covariance matrix is doubled if four times more mocks are used. Overall this means that the number of mocks required to reach the necessary covariance matrix precision for next generation surveys will be much larger than the number currently used.

This presents a bleak picture for the standard method of covariance matrix estimation, in which a delicate balance between the speed, size and accuracy of each simulation must be achieved. Enough simulations must be run to estimate the covariance matrix to high precision, but they must also be large enough to fit the survey and have enough particles. To ease this problem, there have been several recent studies looking at reducing the amount of simulations required to reach a given covariance matrix precision, as well as studies to improve the speed and accuracy of each simulation. An overview of the latter of these was given in Chapter 2.

For a fixed simulation size, one technique for reducing the number of simulations required to reach a given covariance matrix precision is covariance matrix tapering (Kaufman et al., 2008; Paz et al., 2013; Paz & Sanchez, 2015) where the covariance matrix is made more diagonal through the use of a specialised set of tapering functions. As the off-diagonal terms in the covariance matrix generally have low signal-to-noise, diagonalizing the matrix reduces the noise in the inverse covariance matrix and hence decreases the number of simulations needed for model fitting and to estimate parameter likelihoods. Another method, covariance matrix 'shrinkage' Schäfer & Strimmer (2005); Pope & Szapudi (2008); Pearson & Samushia (2015), combines an empirical estimate of the covariance matrix from a small number of mock catalogues with a simple fitting function containing several free parameters. Both of these methods succeed in greatly reducing the number of mock catalogues required to reach a given covariance matrix precision, however they both contain free parameters which must be calibrated. Furthermore, these methods do not overcome the problem that running even a few hundred simulations may be a challenge for next generation surveys.

Rather than just reducing the number of mocks required, this chapter will give a method of reducing the size of each mock catalogue, and hence its computational cost, in a way that still recovers the same covariance matrix had the standard method of fitting the survey into the simulation volume been used. The benefits of this are two-fold. Firstly, rather than reducing the noise in the resultant covariance matrix through methods such as tapering, reducing the simulation volume by, say, a factor of 8 means that each simulation requires 8 times fewer particles and can run much faster. Hence for the same computational cost more simulations can be run and the covariance matrix will be correspondingly more precise. It also means that one no longer has to contend with the (computational)

memory limitations that could pose a problem when trying to simulate the full survey volume. As an added benefit, this provides a solution to the difficulty of even running a single simulation like that in the example used previously.

The key to this method is an understanding of how the covariance matrix scales with the volume of the patch of the universe that is being measured/simulated. In order to investigate this, the first step will be an analytic derivation of the power spectrum estimator and its covariance matrix.

5.2 Analytic Formula for the Power Spectrum

This section details an analytic derivation of the power spectrum, based on the formalism of Feldman et al. (1994). This method is based on the idea that galaxies are Poisson sampled from the underlying density field. As such, the n-point clustering of galaxies can be written in terms of the probability of finding n objects separated by some distance. The use of this method in deriving the two-point correlation function and its Fourier transform, the power spectrum, was briefly shown in Chapter 1. It will be expanded on in this Chapter as a basis upon which to build the analytic derivation of the covariance matrix.

5.2.1 Numerical Conventions

The derivation of the covariance matrix is rather lengthy and for brevity several conventions are adopted both there and in the derivation of the power spectrum estimator. Firstly, the position-dependence of the number density n, weights w, Dirac Delta function δ^D and the connected two-, three- and four-point correlation functions (ξ, ζ, η) will often be only included implicitly or using

$$\begin{aligned} \bar{n}_i = \bar{n}(\boldsymbol{r}_i), \quad w_i = w(\boldsymbol{r}_i), \quad \delta^D_{ij} = \delta^D(\boldsymbol{r}_i - \boldsymbol{r}_j), \\ \xi_{ij} = \xi(|\boldsymbol{r}_i - \boldsymbol{r}_j|), \quad \zeta_{ijk} = \zeta(\boldsymbol{r}_i, \boldsymbol{r}_j, \boldsymbol{r}_k), \quad \eta_{ijkm} = \eta(\boldsymbol{r}_i, \boldsymbol{r}_j, \boldsymbol{r}_k, \boldsymbol{r}_m). \end{aligned} \tag{5.5}$$

Additionally, where it will not be to confusing to do so, the short-hand $\bar{n}_{1234} = \bar{n}_1\bar{n}_2\bar{n}_3\bar{n}_4$ and $w_{1234} = w_1w_2w_3w_4$ will be used. The connected two-, three- and four-point correlation functions form Fourier pairs with the power spectrum, bispectrum and trispectrum respectively. The convention used in this chapter is

$$\xi_{ij} = \int d^3q_{ij} P(q_i)\delta^D(\boldsymbol{q}_{ij}) e^{-i\boldsymbol{q}\cdot\boldsymbol{r}_i - i\boldsymbol{q}_j\cdot\boldsymbol{r}_j}, \tag{5.6}$$

$$\zeta_{ijk} = \int d^3q_{ijk} B(\boldsymbol{q}_i, \boldsymbol{q}_j, \boldsymbol{q}_k)\delta^D(\boldsymbol{q}_{ijk}) e^{-i\boldsymbol{q}_i\cdot\boldsymbol{r}_i - i\boldsymbol{q}_j\cdot\boldsymbol{r}_j - i\boldsymbol{q}_k\cdot\boldsymbol{r}_k}, \tag{5.7}$$

$$\eta_{ijkm} = \int d^3q_{ijkm} T(\boldsymbol{q}_i, \boldsymbol{q}_j, \boldsymbol{q}_k, \boldsymbol{q}_m)\delta^D(\boldsymbol{q}_{ijkm}) e^{-i\boldsymbol{q}_i\cdot\boldsymbol{r}_i - i\boldsymbol{q}_j\cdot\boldsymbol{r}_j - i\boldsymbol{q}_k\cdot\boldsymbol{r}_k - i\boldsymbol{q}_m\cdot\boldsymbol{r}_m}. \tag{5.8}$$

The subscript on d^3q_{ijkm} denotes an integral over independent q-vectors and $\delta^D(\boldsymbol{q}_{ijkm}) = \delta^D(\boldsymbol{q}_i + \boldsymbol{q}_j + \boldsymbol{q}_k + \boldsymbol{q}_m)$ is the Dirac delta function. The Dirac delta function over $\boldsymbol{q}$-vectors will *never* be written δ^D_{ijkm} to avoid unnecessary confusion between Dirac deltas containing $\boldsymbol{r}$-vectors and $\boldsymbol{q}$-vectors. In some instances, for particularly long expressions, the shorthand $T_{ijkm} = T(\boldsymbol{q}_i, \boldsymbol{q}_j, \boldsymbol{q}_k, \boldsymbol{q}_m)$ will also be adopted for the power spectrum, bispectrum and trispectrum.

Finally, as will become apparent later, it is also beneficial to define the quantity

$$G_{p,\ell}(\boldsymbol{k}) = \int d^3r \bar{n}^p(\boldsymbol{r}) w^\ell(\boldsymbol{r}) e^{i\boldsymbol{k}\cdot\boldsymbol{r}}. \tag{5.9}$$

5.2.2 Power Spectrum Estimator

As mentioned in the introduction to this thesis, the starting point for the derivation of the power spectrum covariance matrix is the weighted galaxy overdensity fluctuation, which with the notation adopted in this chapter, is defined as

$$F(\boldsymbol{r}) = \frac{w(\boldsymbol{r})[n^g(\boldsymbol{r}) - \alpha n^s(\boldsymbol{r})]}{(G_{2,2}(0))^{1/2}}. \tag{5.10}$$

This is analogous to Eq.1.92, where the overdensity has been defined in terms of the difference in number densities between the galaxies (n^g) and a 'random', unclustered field (n^s). α is the ratio of galaxies to random points. Fourier transforming and taking the expectation of the square of the modulus results in,

$$\langle |F(\boldsymbol{k})|^2 \rangle = \frac{\int d^3r_{12} w_{12} \langle n_1^g n_2^g - \alpha n_1^g n_2^s - \alpha n_2^g n_1^s + \alpha^2 n_1^s n_2^s \rangle e^{i\boldsymbol{k}\cdot(\boldsymbol{r}_1 - \boldsymbol{r}_2)}}{G_{2,2}(0)}. \tag{5.11}$$

Feldman et al. (1994) provide a useful derivation for the expectation values of the numbers of galaxies and random points in terms of the n-point correlation functions, which is repeated here as the same principles will be of use in deriving the covariance matrix. First consider the expectation value of the integral over some arbitrary function $g(\boldsymbol{r}, \boldsymbol{r}')$,

$$\left\langle \int d^3r_{12} g(\boldsymbol{r}_1, \boldsymbol{r}_2) n_1 n_2 \right\rangle = \int d^3r_{12} g(\boldsymbol{r}_1, \boldsymbol{r}_2) \langle n_1 n_2 \rangle. \tag{5.12}$$

Peebles (1980) present a method of dealing with such integrals by converting them to sums over infinitesimally small cells of value δV that then have an occupation of $n_i = 0$ or 1. This means that $n_i^\gamma = n_i$ The probability of finding an object in cells i and j is given by

$$\langle n_i n_j \rangle = \begin{cases} \bar{n}_i \bar{n}_j \delta V^2 [1 + \xi_{ij}] & i \neq j \\ \bar{n}_i \delta V & i = j \end{cases}. \tag{5.13}$$

Applying this method to the integral in Eq. 5.12

$$\begin{aligned}\left\langle \int d^3r_{12} g(\boldsymbol{r}_1, \boldsymbol{r}_2) n_1 n_2 \right\rangle &= \sum_{i,j} g(\boldsymbol{r}_i, \boldsymbol{r}_j) \langle n_i n_j \rangle \\ &= \sum_{i \neq j} g(\boldsymbol{r}_i, \boldsymbol{r}_j) \bar{n}_i \bar{n}_j \delta V^2 [1 + \xi_{ij}] + \sum_{i=j} g(\boldsymbol{r}_i, \boldsymbol{r}_j) \bar{n}_i \delta V \\ &= \int d^3r_1 \int d^3r_2 g(\boldsymbol{r}_1, \boldsymbol{r}_2) (\bar{n}_1 \bar{n}_2 [1 + \xi_{12}] + \bar{n}_1 \delta^D_{12}).\end{aligned} \tag{5.14}$$

As this expression holds for any arbitrary function $g(\boldsymbol{r}, \boldsymbol{r}')$, then it follows that

$$\langle n^g_1 n^g_2 \rangle = \bar{n}_1 \bar{n}_2 [1 + \xi_{12}] + \bar{n}_1 \delta^D_{12}. \tag{5.15}$$

Similarly, for the other two-point correlators,

$$\langle n^s_1 n^s_2 \rangle = \alpha^{-2} \bar{n}_1 \bar{n}_2 + \alpha^{-1} \bar{n}_1 \delta^D_{12}, \tag{5.16}$$

$$\langle n^g_1 n^s_2 \rangle = \alpha^{-1} \bar{n}_1 \bar{n}_2. \tag{5.17}$$

The physical origin of these terms is as follows: the probability of finding two galaxies at locations r_1 and r_2 is given by the correlation function between the two points, plus the contribution from the Poisson sampling of the field itself which gives an extra factor of the number density when r_1 and r_2 are coincident. The same is true for the random points, except for the corresponding factors of α and the fact that the random points are uncorrelated. As the galaxy and random fields are completely distinct, the two locations are never coincident and there is no contribution from the Poisson sampling. Substituting these into Eq. 5.11 gives

$$\langle |F(\boldsymbol{k})|^2 \rangle = \frac{1}{G_{2,2}(0)} \left[\int d^3r_{12} \bar{n}_{12} w_{12} \xi_{12} e^{i\boldsymbol{k} \cdot (\boldsymbol{r}_1 - \boldsymbol{r}_2)} + (1 + \alpha) \int d^3r_1 \bar{n}_1 w_1^2 \right]. \tag{5.18}$$

Then, substituting the expression for the two-point correlation function in terms of the power spectrum allows one to relate $\langle |F(\boldsymbol{k})|^2 \rangle$ to the power spectrum via

$$\langle |F(\boldsymbol{k})|^2 \rangle = \frac{1}{G_{2,2}(0)} \left[\int d^3q P(q) |G_{1,1}(\boldsymbol{k} - \boldsymbol{q})|^2 + (1 + \alpha) G_{1,2}(0) \right]. \tag{5.19}$$

The last term is the scale-independent 'shot-noise' component arising from the Poisson sampling of the density field. This is such that even for a random unclustered catalogue the measured power spectrum will take the form of a 'white-noise' spectrum with amplitude based on the number density of points. The convolution between $P(q)$ and $|G_{1,1}(\boldsymbol{k} - \boldsymbol{q})|^2$ highlights the effect of the finite volume of the survey. Given an infinite patch in which to measure the power spectrum one can use $|F(\boldsymbol{k})|^2$ to obtain an unbiased estimate of $P(k)$. However, in reality there is some finite size to a survey and in which the power spectrum can be measured. This means that modes larger than the survey are

not captured, leading to what is known as the 'integral constraint'. Furthermore the survey window causes different scales in the power spectrum to become correlated. The net effect of the survey 'window function' is a reduction of power on large scales, and the addition of covariance between scales. As will be shown later, the window function is actually the dominant source of off-diagonal covariance on large scales.

On small scales, however, the window function is expected to have no effect on the power spectrum as all modes are well sampled. As the window function only has support provided that $\boldsymbol{k} - \boldsymbol{q} \approx 0$, an unbiased estimate of the power spectrum on small scales in some bin k_i can be obtained by averaging over all modes in a shell in k-space of volume $V_{\boldsymbol{k}}$,

$$\hat{P}(k) = \int_{V_{\boldsymbol{k}}} \frac{d^3k}{V_{\boldsymbol{k}}} \left[|F(\boldsymbol{k})|^2 - (1+\alpha)\frac{G_{1,2}(0)}{G_{2,2}(0)} \right]. \tag{5.20}$$

This is the FKP estimator for the power spectrum.

5.3 Analytic Formula for the Covariance Matrix

Although the procedure is much longer, requiring one to consider the correlations between four distinct locations, the covariance matrix can also be calculated using a similar method to that used when deriving the FKP power spectrum estimator. The covariance matrix between two scales $\boldsymbol{k}$ and $\boldsymbol{k}'$ is defined as

$$C(\boldsymbol{k}, \boldsymbol{k}') = \langle P(\boldsymbol{k}) P(\boldsymbol{k}') \rangle - \langle P(\boldsymbol{k}) \rangle \langle P(\boldsymbol{k}') \rangle. \tag{5.21}$$

Using the definition of the weighted density field, which has scale-independent shot-noise, which in turn does not contribute to the covariance, this is the same as

$$\begin{aligned} C(\boldsymbol{k}, \boldsymbol{k}') &= \langle |F(\boldsymbol{k})|^2 |F(\boldsymbol{k}')|^2 \rangle - \langle |F(\boldsymbol{k})|^2 \rangle \langle |F(\boldsymbol{k}')|^2 \rangle \\ &= \int d^3 r_{1234} \Big[\langle F_g(\boldsymbol{r}_1) F_g(\boldsymbol{r}_2) F_g(\boldsymbol{r}_3) F_g(\boldsymbol{r}_4) \rangle - \\ &\qquad \langle F_g(\boldsymbol{r}_1) F_g(\boldsymbol{r}_3) \rangle \langle F_g(\boldsymbol{r}_2) F_g(\boldsymbol{r}_4) \rangle \Big] e^{i\boldsymbol{k}\cdot\boldsymbol{r}_1 + i\boldsymbol{k}'\cdot\boldsymbol{r}_2 - i\boldsymbol{k}\cdot\boldsymbol{r}_3 - i\boldsymbol{k}'\cdot\boldsymbol{r}_4}. \end{aligned} \tag{5.22}$$

Eq. 5.10 can be substituted to gain expressions for the two- and four-point functions that will need to be Fourier transformed,

$$\begin{aligned} \langle F_g(\boldsymbol{r}_1) F_g(\boldsymbol{r}_2) F_g(\boldsymbol{r}_3) F_g(\boldsymbol{r}_4) \rangle = \frac{w_{1234}}{(G_{2,2}(0))^2} \Big[&\langle n_1^g n_2^g n_3^g n_4^g - \alpha n_1^g n_2^g n_3^g n_4^s - (3cyc.) \\ &+ \alpha^2 n_1^g n_2^g n_3^s n_4^s + (5cyc.) - \alpha^3 n_1^g n_2^s n_3^s n_4^s - (3cyc.) \\ &+ \alpha^4 n_1^s n_2^s n_3^s n_4^s \rangle \Big], \end{aligned} \tag{5.23}$$

$$\langle F_g(\boldsymbol{r}_1) F_g(\boldsymbol{r}_3) \rangle \langle F_g(\boldsymbol{r}_2) F_g(\boldsymbol{r}_4) \rangle = \frac{w_{1234}}{(G_{2,2}(0))^2} \Big[\langle n_1^g n_3^g - \alpha n_1^g n_3^s - \alpha n_3^g n_1^s + \alpha^2 n_1^s n_3^s \rangle$$

$$\langle n_2^g n_4^g - \alpha n_2^g n_4^s - \alpha n_4^g n_2^s + \alpha^2 n_2^s n_4^s \rangle \Big]. \tag{5.24}$$

The latter expression is simple to calculate and is just the product of the real space version of Eq. 5.11. The other term is a much more complex four-point function. However, the correlators of the form $\langle n_1^g n_2^g n_3^g n_4^g \rangle$ can be solved in a similar way as for the power spectrum derivation, by looking at the probability of finding objects at locations r_1, r_2 etc. in the galaxy and random fields. In the following section a detailed example of the derivation of the term $\langle n_1^g n_2^g n_3^g n_4^g \rangle$ will be given. The same method can be applied to the other correlators, although only the solutions will be given here.

5.3.1 4-point Correlators

The derivation of the expressions for the four-point correlators is analogous to that for the two-point correlators given in Section 5.2.2,

$$\left\langle \int d^3 r_{1234} g(\boldsymbol{r}_1, \boldsymbol{r}_2, \boldsymbol{r}_3, \boldsymbol{r}_4) n_{1234} \right\rangle = \int d^3 r_{1234} g(\boldsymbol{r}_1, \boldsymbol{r}_2, \boldsymbol{r}_3, \boldsymbol{r}_4) \langle n_{1234} \rangle \tag{5.25}$$

Performing the conversion from integrals to sums over infinitesimal cells, there are five distinct types of summation

$$\begin{aligned}
\sum_{i,j,k,m} g(\boldsymbol{r}_i, \boldsymbol{r}_j, \boldsymbol{r}_k, \boldsymbol{r}_m) \langle n_{ijkm} \rangle = & \sum_{i=j=k=m} g(\boldsymbol{r}_i, \boldsymbol{r}_i, \boldsymbol{r}_i, \boldsymbol{r}_i) \langle n_i \rangle \\
& + \sum_{i \neq j=k=m} g(\boldsymbol{r}_i, \boldsymbol{r}_j, \boldsymbol{r}_j, \boldsymbol{r}_j) \langle n_{ij} \rangle + 3cyc. \\
& + \sum_{(i=j) \neq (k=m)} g(\boldsymbol{r}_i, \boldsymbol{r}_i, \boldsymbol{r}_k, \boldsymbol{r}_k) \langle n_{ik} \rangle + 2cyc. \\
& + \sum_{i \neq j \neq k=m} g(\boldsymbol{r}_i, \boldsymbol{r}_j, \boldsymbol{r}_k, \boldsymbol{r}_k) \langle n_{ijk} \rangle + 5cyc. \\
& + \sum_{i \neq j \neq k \neq m} g(\boldsymbol{r}_i, \boldsymbol{r}_j, \boldsymbol{r}_k, \boldsymbol{r}_m) \langle n_{ijkm} \rangle
\end{aligned} \tag{5.26}$$

Again, Peebles (1980) helpfully provides expressions for the expectation values up to the four-point $\langle n_{ijkm} \rangle$. The necessary expressions are of the form

$$\langle n_i \rangle = \bar{n}_i \delta V \tag{5.27}$$

$$\langle n_{ij} \rangle = \bar{n}_{ij} \delta V^2 [1 + \xi_{ij}] \tag{5.28}$$

$$\langle n_{ijk} \rangle = \bar{n}_{ijk} \delta V^3 [1 + \xi_{ij} + \xi_{ik} + \xi_{jk} + \zeta_{ijk}] \tag{5.29}$$

$$\begin{aligned}
\langle n_{ijkm} \rangle = \bar{n}_{ijkm} \delta V^4 [1 + \eta_{ijkm} + \zeta_{ijk} + \zeta_{ijm} + \zeta_{ikm} + \zeta_{jkm} \\
\xi_{ij} + \xi_{ik} + \xi_{im} + \xi_{jk} + \xi_{jm} + \xi_{km} + \xi_{ij}\xi_{km} + \xi_{ik}\xi_{jm} + \xi_{im}\xi_{jk}]
\end{aligned} \tag{5.30}$$

The remaining steps are identical to those of the two-point correlators and will be skipped for brevity. In brief, substituting these, and all necessary permutations, into Eq. 5.26, reverting back to integrals and realising this resultant expression is true for all $g(\boldsymbol{r}_i, \boldsymbol{r}_j, \boldsymbol{r}_k, \boldsymbol{r}_m)$, one arrives at the rather lengthy expression

$$\begin{aligned}
\langle n_1^g n_2^g n_3^g n_4^g \rangle = \bar{n}_{1234} \Bigg\{ & 1 + \eta_{1234} + \zeta_{123} + \zeta_{124} + \zeta_{134} + \zeta_{234} \\
& + \xi_{12} + \xi_{13} + \xi_{14} + \xi_{23} + \xi_{24} + \xi_{34} + \xi_{12}\xi_{34} + \xi_{13}\xi_{24} + \xi_{14}\xi_{23} \\
& + \left(1 + \xi_{12} + \xi_{13} + \xi_{23} + \zeta_{123}\right)\left(\frac{\delta_{14}^D + \delta_{24}^D + \delta_{34}^D}{\bar{n}_4}\right) \\
& + \left(1 + \xi_{12} + \xi_{14} + \xi_{24} + \zeta_{124}\right)\left(\frac{\delta_{13}^D + \delta_{23}^D}{\bar{n}_3}\right) \\
& + \left(1 + \xi_{13} + \xi_{14} + \xi_{34} + \zeta_{134}\right)\left(\frac{\delta_{12}^D}{\bar{n}_2}\right) + \left(1 + \xi_{14}\right)\left(\frac{\delta_{12}^D \delta_{13}^D}{\bar{n}_2 \bar{n}_3}\right) \\
& + \left(1 + \xi_{13}\right)\left(\frac{\delta_{12}^D \delta_{14}^D + \delta_{12}^D \delta_{34}^D + \delta_{14}^D \delta_{23}^D}{\bar{n}_2 \bar{n}_4}\right) \\
& + \left(1 + \xi_{12}\right)\left(\frac{\delta_{13}^D \delta_{14}^D + \delta_{23}^D \delta_{24}^D + \delta_{13}^D \delta_{24}^D}{\bar{n}_3 \bar{n}_4}\right) + \frac{\delta_{12}^D \delta_{13}^D \delta_{14}^D}{\bar{n}_2 \bar{n}_3 \bar{n}_4} \Bigg\}.
\end{aligned} \tag{5.31}$$

Expressions for the four-point correlators involving the galaxy and random fields can be similarly derived. However, a much simpler method is to simply recognise that they follow the form of the above expression but with the necessary two-, three- and four-point correlation functions disappearing as there is no correlation between the galaxy and random fields, or between multiple points on the same random field. The terms containing the Dirac delta function between the galaxy and random fields also disappear as two locations on the separate fields can never be coincident. As such

$$\begin{aligned}
\langle n_1^g n_2^g n_3^g n_4^s \rangle = \alpha^{-1} \bar{n}_{1234} \Bigg\{ & 1 + \xi_{12} + \xi_{13} + \xi_{23} + \zeta_{123} + \left(1 + \xi_{12}\right)\left(\frac{\delta_{13}^D + \delta_{23}^D}{\bar{n}_3}\right) \\
& + \left(1 + \xi_{13}\right)\left(\frac{\delta_{12}^D}{\bar{n}_2}\right) + \frac{\delta_{12}^D \delta_{13}^D}{\bar{n}_2 \bar{n}_3} \Bigg\}
\end{aligned} \tag{5.32}$$

$$\langle n_1^g n_2^g n_3^s n_4^s \rangle = \alpha^{-2} \bar{n}_{1234} \left\{ 1 + \xi_{12} + \frac{\delta_{12}^D}{\bar{n}_2} + \alpha\left(1 + \xi_{12}\right)\frac{\delta_{34}^D}{\bar{n}_4} + \alpha \frac{\delta_{12}^D \delta_{34}^D}{\bar{n}_2 \bar{n}_4} \right\} \tag{5.33}$$

$$\langle n_1^g n_2^s n_3^s n_4^s \rangle = \alpha^{-3} \bar{n}_{1234} \left\{ 1 + \alpha \frac{\delta_{24}^D + \delta_{34}^D}{\bar{n}_4} + \alpha \frac{\delta_{23}^D}{\bar{n}_3} + \alpha^2 \frac{\delta_{23}^D \delta_{24}^D}{\bar{n}_3 \bar{n}_4} \right\} \tag{5.34}$$

$$\begin{aligned}
\langle n_1^s n_2^s n_3^s n_4^s \rangle = \alpha^{-4} \bar{n}_{1234} \Bigg\{ & 1 + \alpha \frac{\delta_{14}^D + \delta_{24}^D + \delta_{34}^D}{\bar{n}_4} + \alpha \frac{\delta_{13}^D + \delta_{23}^D}{\bar{n}_3} + \alpha \frac{\delta_{12}^D}{\bar{n}_2} && (5.35) \\
& + \alpha^2 \frac{\delta_{12}^D \delta_{13}^D}{\bar{n}_2 \bar{n}_3} + \alpha^2 \frac{\delta_{12}^D \delta_{14}^D + \delta_{12}^D \delta_{34}^D + \delta_{14}^D \delta_{23}^D}{\bar{n}_2 \bar{n}_4} && (5.36) \\
& + \alpha^2 \frac{\delta_{13}^D \delta_{14}^D + \delta_{23}^D \delta_{24}^D + \delta_{13}^D \delta_{24}^D}{\bar{n}_3 \bar{n}_4} + \alpha^3 \frac{\delta_{12}^D \delta_{13}^D \delta_{14}^D}{\bar{n}_2 \bar{n}_3 \bar{n}_4} \Bigg\}. && (5.37)
\end{aligned}$$

5.3.2 The Four-point Function

The derivation of the four-point correlators given above is a necessary step in the computation of the power spectrum covariance matrix. With Eqs. 5.31-5.37, and the 10 additional permutations in hand, an expression for four-point function in Eq. 5.23 can be obtained,

$$\begin{aligned}\langle F_g(\boldsymbol{r}_1)F_g(\boldsymbol{r}_2)F_g(\boldsymbol{r}_3)F_g(\boldsymbol{r}_4)\rangle = \frac{\bar{n}_{1234}w_{1234}}{(G_{2,2}(0))^2}\Bigg[&\eta_{1234} + \frac{(1+\alpha^3)\delta^D_{12}\delta^D_{13}\delta^D_{14}}{\bar{n}_2\bar{n}_3\bar{n}_4} \\ &+ \left(\xi_{12} + \frac{(1+\alpha)}{\bar{n}_2}\delta^D_{12}\right)\left(\xi_{34} + \frac{(1+\alpha)}{\bar{n}_4}\delta^D_{34}\right) \\ &+ \left(\xi_{13} + \frac{(1+\alpha)}{\bar{n}_3}\delta^D_{13}\right)\left(\xi_{24} + \frac{(1+\alpha)}{\bar{n}_4}\delta^D_{24}\right) \\ &+ \left(\xi_{14} + \frac{(1+\alpha)}{\bar{n}_4}\delta^D_{14}\right)\left(\xi_{23} + \frac{(1+\alpha)}{\bar{n}_3}\delta^D_{23}\right) \\ &+ \zeta_{123}\frac{\delta^D_{14}}{\bar{n}_4} + \zeta_{123}\frac{\delta^D_{24}}{\bar{n}_4} + \zeta_{123}\frac{\delta^D_{34}}{\bar{n}_4} + \xi_{12}\frac{\delta^D_{13}\delta^D_{14}}{\bar{n}_3\bar{n}_4} + \xi_{12}\frac{\delta^D_{13}\delta^D_{24}}{\bar{n}_3\bar{n}_4} \\ &+ \zeta_{124}\frac{\delta^D_{13}}{\bar{n}_3} + \zeta_{124}\frac{\delta^D_{23}}{\bar{n}_3} + \zeta_{134}\frac{\delta^D_{12}}{\bar{n}_2} + \xi_{12}\frac{\delta^D_{23}\delta^D_{24}}{\bar{n}_3\bar{n}_4} + \xi_{13}\frac{\delta^D_{12}\delta^D_{14}}{\bar{n}_2\bar{n}_4} \\ &+ \xi_{13}\frac{\delta^D_{12}\delta^D_{34}}{\bar{n}_2\bar{n}_4} + \xi_{13}\frac{\delta^D_{14}\delta^D_{23}}{\bar{n}_3\bar{n}_4} + \xi_{14}\frac{\delta^D_{12}\delta^D_{13}}{\bar{n}_2\bar{n}_3}\Bigg].\end{aligned} \tag{5.38}$$

Conveniently, the second term in the expression for the covariance exactly cancels with the fourth term in the above expression

$$\langle F_g(\boldsymbol{r}_1)F_g(\boldsymbol{r}_3)\rangle\langle F_g(\boldsymbol{r}_2)F_g(\boldsymbol{r}_4)\rangle = \frac{\bar{n}_{1234}w_{1234}}{(G_{2,2}(0))^2}\left(\xi_{13} + \frac{(1+\alpha)}{\bar{n}_3}\delta^D_{13}\right)\left(\xi_{24} + \frac{(1+\alpha)}{\bar{n}_4}\delta^D_{24}\right). \tag{5.39}$$

This expression matches the more general derivation found in the Appendices of Smith & Marian (2015).

5.3.3 The Covariance Matrix

An actual estimator for the power spectrum covariance matrix requires one to evaluate the Fourier transform of previous expressions. As with the FKP estimator, the two-, three- and four-point correlation functions can be written in terms of their Fourier counterparts. Making careful use of the Dirac delta function, the different components of the covariance matrix can be written as follows:

Four-point correlation function

$$\int d^3r_{1234}\eta_{1234}\bar{n}_{1234}w_{1234}e^{i\boldsymbol{k}\cdot\boldsymbol{r}_1+i\boldsymbol{k}'\cdot\boldsymbol{r}_2-i\boldsymbol{k}\cdot\boldsymbol{r}_3-i\boldsymbol{k}'\cdot\boldsymbol{r}_4} = \int d^3q_{1234}T_{1234}\delta^D(\boldsymbol{q}_{1234})G_{1,1}(\boldsymbol{k}-\boldsymbol{q_1})G_{1,1}(\boldsymbol{k}'-\boldsymbol{q_2})G^*_{1,1}(\boldsymbol{k}+\boldsymbol{q_3})G^*_{1,1}(\boldsymbol{k}'+\boldsymbol{q_4}) \tag{5.40}$$

Two-point correlation function squared

$$\int d^3r_{1234}\left(\xi_{12}+\frac{(1+\alpha)}{\bar{n}_2}\delta^D_{12}\right)\left(\xi_{34}+\frac{(1+\alpha)}{\bar{n}_4}\delta^D_{34}\right)\bar{n}_{1234}w_{1234}e^{i\boldsymbol{k}\cdot\boldsymbol{r}_1+i\boldsymbol{k}'\cdot\boldsymbol{r}_2-i\boldsymbol{k}\cdot\boldsymbol{r}_3-i\boldsymbol{k}'\cdot\boldsymbol{r}_4} = \int d^3q_{1234}\Big[P_1\delta^D(\boldsymbol{q}_{12})G_{1,1}(\boldsymbol{k}-\boldsymbol{q}_1)G_{1,1}(\boldsymbol{k}'-\boldsymbol{q}_2)+(1+\alpha)G_{1,2}(\boldsymbol{k}+\boldsymbol{k}')\delta^D(\boldsymbol{q}_1)\delta^D(\boldsymbol{q}_2)\Big] \times\Big[P_3\delta^D(\boldsymbol{q}_{34})G^*_{1,1}(\boldsymbol{k}+\boldsymbol{q}_3)G^*_{1,1}(\boldsymbol{k}'+\boldsymbol{q}_4)+(1+\alpha)G^*_{1,2}(\boldsymbol{k}+\boldsymbol{k}')\delta^D(\boldsymbol{q}_3)\delta^D(\boldsymbol{q}_4)\Big] \tag{5.41}$$

The other term of this form, containing the multiplication of two two-point correlation functions, can be similarly derived.

Three-point correlation function

There are six terms containing the three-point correlation function. They all take forms similar to

$$\int d^3r_{1234}\zeta_{124}\frac{\delta^D_{13}}{\bar{n}_3}\bar{n}_{1234}w_{1234}e^{i\boldsymbol{k}\cdot\boldsymbol{r}_1+i\boldsymbol{k}'\cdot\boldsymbol{r}_2-i\boldsymbol{k}\cdot\boldsymbol{r}_3-i\boldsymbol{k}'\cdot\boldsymbol{r}_4} = \int d^3q_{1234}B_{124}\delta^D(\boldsymbol{q}_{124})\delta^D(\boldsymbol{q}_3)G^*_{1,2}(\boldsymbol{q}_1)G_{1,1}(\boldsymbol{k}'-\boldsymbol{q}_2)G^*_{1,1}(\boldsymbol{k}'+\boldsymbol{q}_4) \tag{5.42}$$

Two-point correlation function terms

There are, in total, seven terms in the real space covariance matrix containing a single two-point correlation function and two Dirac delta functions. The procedure for calculating these is much the same as for the three-point correlation function terms. For example,

$$\int d^3r_{1234}\xi_{12}\frac{\delta^D_{13}\delta^D_{24}}{\bar{n}_3\bar{n}_4}\bar{n}_{1234}w_{1234}e^{i\boldsymbol{k}\cdot\boldsymbol{r}_1+i\boldsymbol{k}'\cdot\boldsymbol{r}_2-i\boldsymbol{k}\cdot\boldsymbol{r}_3-i\boldsymbol{k}'\cdot\boldsymbol{r}_4} = \int d^3q_{1234}P_1\delta^D(\boldsymbol{q}_{12})\delta^D(\boldsymbol{q}_3)\delta^D(\boldsymbol{q}_4)G^*_{1,2}(\boldsymbol{q}_1)G^*_{1,2}(\boldsymbol{q}_2) \tag{5.43}$$

Constant term

The final remaining term does not contain any correlation functions and consists of only a constant and a trio of Dirac delta functions,

$$\int d^3r_{1234}(1+\alpha^3)\frac{\delta^D_{12}\delta^D_{13}\delta^D_{14}}{\bar{n}_2\bar{n}_3\bar{n}_4}\bar{n}_{1234}w_{1234}e^{i\boldsymbol{k}\cdot\boldsymbol{r}_1+i\boldsymbol{k}'\cdot\boldsymbol{r}_2-i\boldsymbol{k}\cdot\boldsymbol{r}_3-i\boldsymbol{k}'\cdot\boldsymbol{r}_4} = (1+\alpha^3)G_{1,4}(0) \tag{5.44}$$

Combined Terms

Using the above expressions and applying the same techniques to derive solutions for all the remaining terms in Eq. 5.38, one arrives at the full expression for the k-space covariance matrix,

$$\begin{aligned}
C(\boldsymbol{k},\boldsymbol{k}') = \frac{1}{(G_{2,2}(0))^2}\int d^3q_{1234}\Bigg[& \\
& T_{1234}\delta^D(\boldsymbol{q}_{1234})G_{1,1}(\boldsymbol{k}-\boldsymbol{q}_1)G_{1,1}(\boldsymbol{k}'-\boldsymbol{q}_2)G^*_{1,1}(\boldsymbol{k}+\boldsymbol{q}_3)G^*_{1,1}(\boldsymbol{k}'+\boldsymbol{q}_4) \\
& + \Big(P_1\delta^D(\boldsymbol{q}_{12})G_{1,1}(\boldsymbol{k}-\boldsymbol{q}_1)G_{1,1}(\boldsymbol{k}'-\boldsymbol{q}_2) + (1+\alpha)G_{1,2}(\boldsymbol{k}+\boldsymbol{k}')\delta^D(\boldsymbol{q}_1)\delta^D(\boldsymbol{q}_2)\Big) \\
& \times \Big(P_3\delta^D(\boldsymbol{q}_{34})G^*_{1,1}(\boldsymbol{k}+\boldsymbol{q}_3)G^*_{1,1}(\boldsymbol{k}'+\boldsymbol{q}_4) + (1+\alpha)G^*_{1,2}(\boldsymbol{k}+\boldsymbol{k}')\delta^D(\boldsymbol{q}_3)\delta^D(\boldsymbol{q}_4)\Big) \\
& + \Big(P_1\delta^D(\boldsymbol{q}_{14})G_{1,1}(\boldsymbol{k}-\boldsymbol{q}_1)G^*_{1,1}(\boldsymbol{k}'+\boldsymbol{q}_4) + (1+\alpha)G_{1,2}(\boldsymbol{k}-\boldsymbol{k}')\delta^D(\boldsymbol{q}_1)\delta^D(\boldsymbol{q}_4)\Big) \\
& \times \Big(P_2\delta^D(\boldsymbol{q}_{23})G_{1,1}(\boldsymbol{k}'-\boldsymbol{q}_2)G^*_{1,1}(\boldsymbol{k}+\boldsymbol{q}_3) + (1+\alpha)G_{1,2}(\boldsymbol{k}-\boldsymbol{k}')\delta^D(\boldsymbol{q}_2)\delta^D(\boldsymbol{q}_3)\Big) \\
& + B_{134}\delta^D(\boldsymbol{q}_{134})\delta^D(\boldsymbol{q}_2)G_{1,2}(\boldsymbol{k}+\boldsymbol{k}'-\boldsymbol{q}_1)G^*_{1,1}(\boldsymbol{k}+\boldsymbol{q}_3)G^*_{1,1}(\boldsymbol{k}'+\boldsymbol{q}_4) \\
& + B_{124}\delta^D(\boldsymbol{q}_{124})\delta^D(\boldsymbol{q}_3)G^*_{1,2}(\boldsymbol{q}_1)G_{1,1}(\boldsymbol{k}'-\boldsymbol{q}_2)G^*_{1,1}(\boldsymbol{k}'+\boldsymbol{q}_4) \\
& + B_{124}\delta^D(\boldsymbol{q}_{124})\delta^D(\boldsymbol{q}_3)G_{1,2}(\boldsymbol{k}'-\boldsymbol{k}-\boldsymbol{q}_2)G_{1,1}(\boldsymbol{k}-\boldsymbol{q}_1)G^*_{1,1}(\boldsymbol{k}'+\boldsymbol{q}_4) \\
& + B_{123}\delta^D(\boldsymbol{q}_{123})\delta^D(\boldsymbol{q}_4)G_{1,2}(\boldsymbol{k}-\boldsymbol{k}'-\boldsymbol{q}_1)G_{1,1}(\boldsymbol{k}'-\boldsymbol{q}_2)G^*_{1,1}(\boldsymbol{k}+\boldsymbol{q}_3) \\
& + B_{123}\delta^D(\boldsymbol{q}_{123})\delta^D(\boldsymbol{q}_4)G^*_{1,2}(\boldsymbol{q}_2)G_{1,1}(\boldsymbol{k}-\boldsymbol{q}_1)G^*_{1,1}(\boldsymbol{k}+\boldsymbol{q}_3) \\
& + B_{123}\delta^D(\boldsymbol{q}_{123})\delta^D(\boldsymbol{q}_4)G^*_{1,2}(\boldsymbol{k}+\boldsymbol{k}'+\boldsymbol{q}_3)G_{1,1}(\boldsymbol{k}-\boldsymbol{q}_1)G_{1,1}(\boldsymbol{k}'-\boldsymbol{q}_2) \\
& + P_1\delta^D(\boldsymbol{q}_{12})\delta^D(\boldsymbol{q}_3)\delta^D(\boldsymbol{q}_4)G^*_{1,3}(\boldsymbol{k}'+\boldsymbol{q}_1)G_{1,1}(\boldsymbol{k}'-\boldsymbol{q}_2) \\
& + P_1\delta^D(\boldsymbol{q}_{12})\delta^D(\boldsymbol{q}_3)\delta^D(\boldsymbol{q}_4)G^*_{1,2}(\boldsymbol{q}_1)G^*_{1,2}(\boldsymbol{q}_2) \\
& + P_1\delta^D(\boldsymbol{q}_{12})\delta^D(\boldsymbol{q}_3)\delta^D(\boldsymbol{q}_4)G_{1,2}(\boldsymbol{k}-\boldsymbol{k}'-\boldsymbol{q}_1)G_{1,2}(\boldsymbol{k}'-\boldsymbol{k}-\boldsymbol{q}_2) \\
& + P_1\delta^D(\boldsymbol{q}_{12})\delta^D(\boldsymbol{q}_3)\delta^D(\boldsymbol{q}_4)G_{1,1}(\boldsymbol{k}-\boldsymbol{q}_1)G^*_{1,3}(\boldsymbol{k}+\boldsymbol{q}_2) \\
& + P_1\delta^D(\boldsymbol{q}_{13})\delta^D(\boldsymbol{q}_2)\delta^D(\boldsymbol{q}_4)G^*_{1,1}(\boldsymbol{k}+\boldsymbol{q}_3)G_{1,3}(\boldsymbol{k}-\boldsymbol{q}_1) \\
& + P_1\delta^D(\boldsymbol{q}_{13})\delta^D(\boldsymbol{q}_2)\delta^D(\boldsymbol{q}_4)G_{1,2}(\boldsymbol{k}+\boldsymbol{k}'-\boldsymbol{q}_1)G^*_{1,2}(\boldsymbol{k}+\boldsymbol{k}'+\boldsymbol{q}_3) \\
& + P_1\delta^D(\boldsymbol{q}_{14})\delta^D(\boldsymbol{q}_2)\delta^D(\boldsymbol{q}_3)G_{1,3}(\boldsymbol{k}'-\boldsymbol{q}_1)G^*_{1,1}(\boldsymbol{k}'+\boldsymbol{q}_4) \\
& + (1+\alpha^3)G_{1,4}(0)\delta^D(\boldsymbol{q}_1)\delta^D(\boldsymbol{q}_2)\delta^D(\boldsymbol{q}_3)\delta^D(\boldsymbol{q}_4)\Bigg].
\end{aligned} \tag{5.45}$$

This expression can be reduced slightly if one makes use of the symmetry of the $\boldsymbol{k}$-vectors which allows $\boldsymbol{k}' \to -\boldsymbol{k}'$ and $\boldsymbol{k} \to -\boldsymbol{k}$ and by simply renaming some of the indices, which can be done freely as the integral can be applied to each term independently. In this case,

$$
\begin{aligned}
C(\boldsymbol{k},\boldsymbol{k}') = 2\Bigg(& \int d^3q_1 P_1 \frac{G_{1,1}(\boldsymbol{k}-\boldsymbol{q}_1)G_{1,1}(\boldsymbol{k}'+\boldsymbol{q}_1)}{G_{2,2}(0)} + (1+\alpha)\frac{G_{1,2}(\boldsymbol{k}+\boldsymbol{k}')}{G_{2,2}(0)}\Bigg) \\
& \times \left(\int d^3q_2 P_2 \frac{G^*_{1,1}(\boldsymbol{k}-\boldsymbol{q}_2)G^*_{1,1}(\boldsymbol{k}'+\boldsymbol{q}_2)}{G_{2,2}(0)} + (1+\alpha)\frac{G^*_{1,2}(\boldsymbol{k}+\boldsymbol{k}')}{G_{2,2}(0)}\right) \\
& + \int d^3q_{1234} T_{1234}\delta^D(\boldsymbol{q}_{1234}) \frac{G_{1,1}(\boldsymbol{k}-\boldsymbol{q}_1)G_{1,1}(\boldsymbol{k}'-\boldsymbol{q}_2)G^*_{1,1}(\boldsymbol{k}+\boldsymbol{q}_3)G^*_{1,1}(\boldsymbol{k}'+\boldsymbol{q}_4)}{(G_{2,2}(0))^2} \\
& + 4\int d^3q_{123} B_{123}\delta^D(\boldsymbol{q}_{123}) \frac{G_{1,2}(\boldsymbol{k}+\boldsymbol{k}'-\boldsymbol{q}_1)G^*_{1,1}(\boldsymbol{k}+\boldsymbol{q}_2)G^*_{1,1}(\boldsymbol{k}'+\boldsymbol{q}_3)}{(G_{2,2}(0))^2} \\
& + \int d^3q_{123} B_{123}\delta^D(\boldsymbol{q}_{123}) \frac{G^*_{1,2}(\boldsymbol{q}_1)G_{1,1}(\boldsymbol{k}'-\boldsymbol{q}_2)G^*_{1,1}(\boldsymbol{k}'+\boldsymbol{q}_3)}{(G_{2,2}(0))^2} \\
& + \int d^3q_{123} B_{123}\delta^D(\boldsymbol{q}_{123}) \frac{G^*_{1,2}(\boldsymbol{q}_1)G_{1,1}(\boldsymbol{k}-\boldsymbol{q}_2)G^*_{1,1}(\boldsymbol{k}'+\boldsymbol{q}_3)}{(G_{2,2}(0))^2} \\
& + 2\int d^3q_1 P_1 \left(\frac{G_{1,3}(\boldsymbol{k}-\boldsymbol{q}_1)G^*_{1,1}(\boldsymbol{k}-\boldsymbol{q}_1) + G_{1,3}(\boldsymbol{k}'-\boldsymbol{q}_1)G^*_{1,1}(\boldsymbol{k}'-\boldsymbol{q}_1)}{(G_{2,2}(0))^2}\right) \\
& + 2\int d^3q_1 P_1 \frac{|G_{1,2}(\boldsymbol{k}+\boldsymbol{k}'-\boldsymbol{q}_1)|^2}{(G_{2,2}(0))^2} + \int d^3q_1 P_1 \frac{|G_{1,2}(\boldsymbol{q}_1)|^2}{(G_{2,2}(0))^2} \\
& + (1+\alpha^3)\frac{G_{1,4}(0)}{(G_{2,2}(0))^2}\Bigg].
\end{aligned}
\tag{5.46}
$$

This expression consists of several well known terms. The first is the dominant component of the covariance matrix on both large and small scales for any realistic survey. This consists of the combination of power spectrum squared and shot-noise that has been well established in the literature (Feldman et al., 1994; Tegmark, 1997). On large scales the power spectrum itself, and hence the volume available to measure this in, dominates the covariance matrix. This is the cosmic variance limit. On small scales the shot-noise term (containing $G_{1,2}(\boldsymbol{k}+\boldsymbol{k}')$) becomes larger than the power spectrum and so the covariance is limited by the number of tracers available for measuring the power spectrum.

As will be shown in the next section, in the absence of a window function this first term only contributes to diagonal terms in the covariance matrix. However the presence of a window function causes a convolution with the power spectrum which then introduces off-diagonal covariance. Indeed the effect of the window function makes this component the largest source of off-diagonal covariance on large scales. On small scales, where the effect of the window function becomes small however the other terms can start to dominate the off-diagonal covariance.

The primary of these is the trispectrum term, arising from the connected four-point function within the covariance matrix. In the absence of shot-noise or a window function, this contribution still remains, and indeed comes to dominate the diagonal covariance on small scales. The nature of the trispectrum is such that even if there is no shot noise or window function there is still off-diagonal covariance, arising only from the trispectrum. For a Gaussian random field, where there is no trispectrum, there is no small-scale off-diagonal covariance. In the presence of shot-noise, the contribution to the diagonal covariance from the trispectrum becomes generally subdominant, except on very small scales. Similarly, in the presence of a window function the trispectrum becomes less important in the off-diagonal regime, except on small scales. The survey window itself does convolve the trispectrum, affecting quadrilaterals with side lengths of order the survey volume. This mode-coupling introduced by the window function causes the long wavelength trispectrum to couple with the small wavelength modes, introducing additional small scale covariance. This additional covariance is called *supersample* covariance and will be presented in greater detail later.

The final contributions to the covariance matrix are higher order shot noise terms which only come into play on very small scales and are subdominant in comparison to the terms described above. In the absence of shot-noise these all vanish. There are three types. The first contains contributions from the three-point clustering and the shot-noise. These are of the order of the bispectrum multiplied by $\bar{n}^{-1}$. The $\bar{n}$ dependence can be inferred from the ratio of $G_{1,2}$ terms to $G_{2,2}$ terms. As the bispectrum is small for fields close to Gaussian, this will have a lower impact on the covariance than the shot-noise terms detailed previously. The second set of higher order shot-noise components is proportional to the power spectrum and on the order of $\bar{n}^{-2}$, whilst the last, constant term is on the order of $\bar{n}^{-3}$. Hence these remain negligible far into the non-linear regime.

5.4 Covariance Matrix in the Small-Scale Limit.

Several studies have looked at the power spectrum covariance matrix in the absence of a window function and as would be measured in a simulated periodic box. Both Meiksin & White (1999) and Scoccimarro et al. (1999) derived the full expressions for the covariance from such a simulation, whilst Feldman et al. (1994) and Tegmark (1997) give expressions in the small-scale, Gaussian limit, neglecting higher order shot-noise contributions. On small scales, where the convolution with the window function is negligible, and in the absence of position dependent number densities and weights (i.e., in a dark matter simulation), it should be expected that these previous results are recovered from Eq. 5.46. This serves as a useful consistency check for the previous derivation.

The first step in this procedure is to calculate the small-scale limit of Eq. 5.46. The

small scale limit of each term in the full expression for the covariance matrix can be derived by making transformations of the form $\boldsymbol{q}' = \boldsymbol{k} - \boldsymbol{q}$ and then taking the limit that the window function only has an effect on the largest scales. Mathematically, this translates to a window function that only has support for $\boldsymbol{q}' \approx 0$. This in turn leads to $\boldsymbol{k} - \boldsymbol{q}' \approx \boldsymbol{k}$. As such, the power spectrum, bispectrum and trispectrum become functions of only the vectors $\boldsymbol{k}$ and $\boldsymbol{k}'$ and can be removed from the convolution with the window function.

Applying these steps to some of the individual components of Eq. 5.46 results in the following examples:

Trispectrum Term

$$\begin{aligned}
&\int d^3q_{1234} T_{1234} \delta^D(\boldsymbol{q}_{1234}) G_{1,1}(\boldsymbol{k} - \boldsymbol{q_1}) G_{1,1}(\boldsymbol{k}' - \boldsymbol{q_2}) G^*_{1,1}(\boldsymbol{k} + \boldsymbol{q_3}) G^*_{1,1}(\boldsymbol{k}' + \boldsymbol{q_4}) \\
&= \frac{T(\boldsymbol{k}, \boldsymbol{k}', -\boldsymbol{k}, -\boldsymbol{k}')}{(G_{2,2}(0))^2} \int d^3q'_{1234} \delta(\boldsymbol{q}'_{1234}) G_{1,1}(\boldsymbol{q}'_1) G_{1,1}(\boldsymbol{q}'_2) G_{1,1}(\boldsymbol{q}'_3) G_{1,1}(\boldsymbol{q}'_3) \\
&= T(\boldsymbol{k}, \boldsymbol{k}', -\boldsymbol{k}, -\boldsymbol{k}') \frac{G_{4,4}(0)}{(G_{2,2}(0))^2}
\end{aligned} \tag{5.47}$$

Power Spectrum Squared Term

$$\begin{aligned}
&\left(\int d^3q_1 P_1 \frac{G_{1,1}(\boldsymbol{k} - \boldsymbol{q}_1) G_{1,1}(\boldsymbol{k}' + \boldsymbol{q}_1)}{G_{2,2}(0)} + (1+\alpha) \frac{G_{1,2}(\boldsymbol{k} + \boldsymbol{k}')}{G_{2,2}(0)} \right) \\
&\qquad \times \left(\int d^3q_2 P_2 \frac{G^*_{1,1}(\boldsymbol{k} - \boldsymbol{q}_2) G^*_{1,1}(\boldsymbol{k}' + \boldsymbol{q}_2)}{G_{2,2}(0)} + (1+\alpha) \frac{G^*_{1,2}(\boldsymbol{k} + \boldsymbol{k}')}{G_{2,2}(0)} \right) \\
&= \frac{P^2(k)}{(G_{2,2}(0))^2} \left| G_{2,2}(\boldsymbol{k} + \boldsymbol{k}') + \frac{(1+\alpha)}{P(k)} G_{1,2}(\boldsymbol{k} + \boldsymbol{k}') \right|^2
\end{aligned} \tag{5.48}$$

Bispectrum Term

$$\begin{aligned}
&\int d^3q_{123} B_{123} \delta^D(\boldsymbol{q}_{123}) \frac{G_{1,2}(\boldsymbol{k} + \boldsymbol{k}' - \boldsymbol{q}_1) G^*_{1,1}(\boldsymbol{k} + \boldsymbol{q}_2) G^*_{1,1}(\boldsymbol{k}' + \boldsymbol{q}_3)}{(G_{2,2}(0))^2} \\
&\qquad = B(\boldsymbol{k} + \boldsymbol{k}', -\boldsymbol{k}, -\boldsymbol{k}') \frac{G_{3,4}(0)}{(G_{2,2}(0))^2}
\end{aligned} \tag{5.49}$$

Power Spectrum Term

$$\begin{aligned}
&\int d^3q_1 P_1 \left(\frac{G_{1,3}(\boldsymbol{k} - \boldsymbol{q}_1) G^*_{1,1}(\boldsymbol{k} - \boldsymbol{q}_1) + G_{1,3}(\boldsymbol{k} - \boldsymbol{q}_1) G^*_{1,1}(\boldsymbol{k} - \boldsymbol{q}_1)}{(G_{2,2}(0))^2} \right) \\
&\qquad = \Big(P(k) + P(k') \Big) \frac{G_{2,4}(0)}{(G_{2,2}(0))^2}
\end{aligned} \tag{5.50}$$

Combined Terms

Calculating the small-scale limit of every term in the covariance matrix results in the small-scale covariance,

$$\begin{aligned} C^{ss}(\boldsymbol{k},\boldsymbol{k}') = 2\frac{P^2(k)}{(G_{2,2}(0))^2}\left|G_{2,2}(\boldsymbol{k}+\boldsymbol{k}') + \frac{(1+\alpha)}{P(k)}G_{1,2}(\boldsymbol{k}+\boldsymbol{k}')\right|^2 \\ + T(\boldsymbol{k},\boldsymbol{k}',-\boldsymbol{k},-\boldsymbol{k}')\frac{G_{4,4}(0)}{(G_{2,2}(0))^2} + (1+\alpha^3)\frac{G_{1,4}(0)}{(G_{2,2}(0))^2} \\ + \left(4B(\boldsymbol{k}+\boldsymbol{k}',-\boldsymbol{k},-\boldsymbol{k}') + B(0,\boldsymbol{k}',-\boldsymbol{k}') + B(\boldsymbol{k},0,-\boldsymbol{k})\right)\frac{G_{3,4}(0)}{(G_{2,2}(0))^2} \\ + \left(2P(k') + 2P(k) + P(|\boldsymbol{k}-\boldsymbol{k}'|) + P(|\boldsymbol{k}+\boldsymbol{k}'|)\right)\frac{G_{2,4}(0)}{(G_{2,2}(0))^2}. \end{aligned} \tag{5.51}$$

The final result in the FKP derivation the power spectrum estimator was the bin-averaged power spectrum, i.e., the power spectrum averaged over some volume in k-space. The true quantity of interest then in the bin-averaged covariance matrix, averaged over two shells in k-space centred on k_i and k_j,

$$C(k_i,k_j) = \int_{V_{\boldsymbol{k}_i}}\frac{d^3k}{V_{\boldsymbol{k}_i}}\int_{V_{\boldsymbol{k}_j}}\frac{d^3k'}{V_{\boldsymbol{k}_j}}C(\boldsymbol{k},\boldsymbol{k}'). \tag{5.52}$$

This is what will be compared to previous studies. The bin averaged covariance can be calculated by bin-averaging each component separately. Most of the terms are trivial to determine if one relates them to the bin-averaged power spectrum, bispectrum and trispectrum, however the first term in Eq. 5.51 requires extra care. This term is the dominant term in the covariance matrix, and the only term if one neglects non-Gaussian and higher order shot noise components. As such bin-averaging this should recover the expressions found in Feldman et al. (1994) and Tegmark (1997).

Multiplying out this term gives

$$\begin{aligned} &\int_{V_{\boldsymbol{k}_i}}\frac{d^3k}{V_{\boldsymbol{k}_i}}\int_{V_{\boldsymbol{k}_j}}\frac{d^3k'}{V_{\boldsymbol{k}_j}}\frac{P^2(k)}{(G_{2,2}(0))^2}\left|G_{2,2}(\boldsymbol{k}+\boldsymbol{k}') + \frac{(1+\alpha)}{P(k)}G_{1,2}(\boldsymbol{k}+\boldsymbol{k}')\right|^2 \\ &= \int_{V_{\boldsymbol{k}_i}}\frac{d^3k}{V_{\boldsymbol{k}_i}}\int_{V_{\boldsymbol{k}_j}}\frac{d^3k'}{V_{\boldsymbol{k}_j}}P^2(k)\frac{|G_{2,2}(\boldsymbol{k}+\boldsymbol{k}')|^2}{(G_{2,2}(0))^2} \\ &+ (1+\alpha)\int_{V_{\boldsymbol{k}_i}}\frac{d^3k}{V_{\boldsymbol{k}_i}}\int_{V_{\boldsymbol{k}_j}}\frac{d^3k'}{V_{\boldsymbol{k}_j}}P(k)\frac{G_{2,2}(\boldsymbol{k}+\boldsymbol{k}')G^*_{1,2}(\boldsymbol{k}+\boldsymbol{k}')}{(G_{2,2}(0))^2} \\ &+ (1+\alpha)\int_{V_{\boldsymbol{k}_i}}\frac{d^3k}{V_{\boldsymbol{k}_i}}\int_{V_{\boldsymbol{k}_j}}\frac{d^3k'}{V_{\boldsymbol{k}_j}}P(k)\frac{G^*_{2,2}(\boldsymbol{k}+\boldsymbol{k}')G_{1,2}(\boldsymbol{k}+\boldsymbol{k}')}{(G_{2,2}(0))^2} \\ &+ (1+\alpha)^2\int_{V_{\boldsymbol{k}_i}}\frac{d^3k}{V_{\boldsymbol{k}_i}}\int_{V_{\boldsymbol{k}_j}}\frac{d^3k'}{V_{\boldsymbol{k}_j}}\frac{|G_{1,2}(\boldsymbol{k}+\boldsymbol{k}')|^2}{(G_{2,2}(0))^2} \end{aligned} \quad . \tag{5.53}$$

Solving each of these four components uses the same method, which is adopted from that found in the Appendices of Smith & Marian (2015). Looking only at the first term,

for sufficiently narrow measurement bins, the power spectrum can be removed from the integral. This is because it is expected that the coherence length of the power spectrum, the scale over which it varies, is much larger than the bin width. The power spectrum in turn becomes the bin-averaged power spectrum from Eq. 5.20,

$$\int_{V_{\boldsymbol{k}_i}} \frac{d^3k}{V_{\boldsymbol{k}_i}} \int_{V_{\boldsymbol{k}_j}} \frac{d^3k'}{V_{\boldsymbol{k}_j}} P^2(k) \frac{|G_{2,2}(\boldsymbol{k}+\boldsymbol{k}')|^2}{(G_{2,2}(0))^2} = P^2(k_i) \int_{V_{\boldsymbol{k}_i}} \frac{d^3k}{V_{\boldsymbol{k}_i}} \int_{V_{\boldsymbol{k}_j}} \frac{d^3k'}{V_{\boldsymbol{k}_j}} \frac{|G_{2,2}(\boldsymbol{k}+\boldsymbol{k}')|^2}{(G_{2,2}(0))^2}. \tag{5.54}$$

Writing the $|G_{2,2}(\boldsymbol{k}+\boldsymbol{k}')|^2$ term in its integral form and transforming coordinates via $r' \to r_2 - r_1$ results in

$$\begin{aligned} &P^2(k_i) \int_{V_{\boldsymbol{k}_i}} \frac{d^3k}{V_{\boldsymbol{k}_i}} \int_{V_{\boldsymbol{k}_j}} \frac{d^3k'}{V_{\boldsymbol{k}_j}} \frac{|G_{2,2}(\boldsymbol{k}+\boldsymbol{k}')|^2}{(G_{2,2}(0))^2} \\ &\quad = \frac{P^2(k_i)}{(G_{2,2}(0))^2} \int d^3r_1 \bar{n}_1^2 w_1^2 \int d^3r_2 \bar{n}_2^2 w_2^2 \int_{V_{\boldsymbol{k}_i}} \frac{d^3k}{V_{\boldsymbol{k}_i}} e^{i\boldsymbol{k}\cdot(\boldsymbol{r}_1-\boldsymbol{r}_2)} \int_{V_{\boldsymbol{k}_j}} \frac{d^3k'}{V_{\boldsymbol{k}_j}} e^{i\boldsymbol{k}'\cdot(\boldsymbol{r}_1-\boldsymbol{r}_2)} \\ &\quad = \frac{P^2(k_i)}{(G_{2,2}(0))^2} \int d^3r_1 \bar{n}_1^2 w_1^2 \int d^3r' \bar{n}^2(\boldsymbol{r}'+\boldsymbol{r}_1) w^2(\boldsymbol{r}'+\boldsymbol{r}_1) \int_{V_{\boldsymbol{k}_i}} \frac{d^3k}{V_{\boldsymbol{k}_i}} e^{-i\boldsymbol{k}\cdot\boldsymbol{r}'} \int_{V_{\boldsymbol{k}_j}} \frac{d^3k'}{V_{\boldsymbol{k}_j}} e^{-i\boldsymbol{k}'\cdot\boldsymbol{r}'}. \end{aligned} \tag{5.55}$$

Here the properties of the window function on small scales can be used again. As the survey volume is much larger than the scales of interest, the window function is not expected to vary much as a function of r'. Hence a reasonable approximation in this regime is $\bar{n}(\boldsymbol{r}'+\boldsymbol{r}_1) \approx \bar{n}(\boldsymbol{r}_1)$. Utilising this gives the final expression

$$P^2(k_i) \int_{V_{\boldsymbol{k}_i}} \frac{d^3k}{V_{\boldsymbol{k}_i}} \int_{V_{\boldsymbol{k}_j}} \frac{d^3k'}{V_{\boldsymbol{k}_j}} \frac{|G_{2,2}(\boldsymbol{k}+\boldsymbol{k}')|^2}{(G_{2,2}(0))^2} = \frac{(2\pi)^3 P^2(k_i)}{V_{k_i}} \frac{G_{4,4}(0)}{(G_{2,2}(0))^2} \delta^D(k_i - k_j). \tag{5.56}$$

Similar expressions are obtained for the other three parts of Eq. 5.53. Overall

$$\begin{aligned} &\int_{V_{\boldsymbol{k}_i}} \frac{d^3k}{V_{\boldsymbol{k}_i}} \int_{V_{\boldsymbol{k}_j}} \frac{d^3k'}{V_{\boldsymbol{k}_j}} \frac{P^2(k)}{(G_{2,2}(0))^2} \left| G_{2,2}(\boldsymbol{k}+\boldsymbol{k}') + \frac{(1+\alpha)}{P(k)} G_{1,2}(\boldsymbol{k}+\boldsymbol{k}') \right|^2 \\ &= \frac{(2\pi)^3 P^2(k) \delta^D(k_i - k_j)}{V_{k_i} (G_{2,2}(0))^2} \left(G_{4,4}(0) + 2\frac{(1+\alpha)}{P(k_i)} G_{3,4}(0) + \frac{(1+\alpha)^2}{P^2(k_i)} G_{2,4}(0) \right) \\ &= \frac{(2\pi)^3 P^2(k) \delta^D(k_i - k_j)}{V_{k_i} V_{\mathrm{eff}(k_i)}} \end{aligned} \tag{5.57}$$

where

$$V_{\mathrm{eff}}(k_i) = \frac{(G_{2,2}(0))^2}{\int d^3r \bar{n}^4(\boldsymbol{r}) w^4(\boldsymbol{r}) \left(1 + \frac{1+\alpha}{\bar{n}(\boldsymbol{r}) P(k_i)} \right)^2}. \tag{5.58}$$

This expression matches the well known expressions for the covariance matrix from Feldman et al. (1994) and Tegmark (1997).

Incorporating this into the full expression for the small-scale covariance matrix, in-

cluding non-Gaussian and higher order shot-noise terms, one arrives at

$$\begin{aligned} C^{ss}(k_i,k_j) = \bar{T}(k_i,k_j)\frac{G_{4,4}(0)}{(G_{2,2}(0))^2} + 2\frac{(2\pi)^3\bar{P}^2(k_i)\delta^D(k_i-k_j)}{V_{k_i}V_{\rm eff}(k_i)} \\ + \left(4\bar{B}(k_i,k_j)+\bar{B}(0,k_j)+\bar{B}(k_i,0)\right)\frac{G_{3,4}(0)}{(G_{2,2}(0))^2} \\ + 2\left(\bar{P}(k_i)+\bar{P}(k_j)+\bar{P}(k_i,k_j)\right)\frac{G_{2,4}(0)}{(G_{2,2}(0))^2} \\ + (1+\alpha^3)\frac{G_{1,4}(0)}{(G_{2,2}(0))^2}. \end{aligned} \tag{5.59}$$

where, on top of the previous definitions,

$$\bar{T}(k_i,k_j) = \int_{V_{k_i}}\frac{d^3k}{V_{k_i}}\int_{V_{k_j}}\frac{d^3k'}{V_{k_j}}T(\boldsymbol{k},\boldsymbol{k'},-\boldsymbol{k},-\boldsymbol{k'}), \tag{5.60}$$

$$\bar{B}(k_i,k_j) = \int_{V_{k_i}}\frac{d^3k}{V_{k_i}}\int_{V_{k_j}}\frac{d^3k'}{V_{k_j}}B(\boldsymbol{k},\boldsymbol{k'},-\boldsymbol{k}-\boldsymbol{k'}) \quad \text{and} \tag{5.61}$$

$$\bar{P}(k_i,k_j) = \int_{V_{k_i}}\frac{d^3k}{V_{k_i}}\int_{V_{k_j}}\frac{d^3k'}{V_{k_j}}P(|\boldsymbol{k}-\boldsymbol{k'}|) = \int_{V_{k_i}}\frac{d^3k}{V_{k_i}}\int_{V_{k_j}}\frac{d^3k'}{V_{k_j}}P(|\boldsymbol{k}+\boldsymbol{k'}|). \tag{5.62}$$

5.5 Covariance Matrix with no Window Function

In the absence of a window function the covariance matrix on all scales takes the form of the small scale covariance, so that $C(\boldsymbol{k},\boldsymbol{k'}) = C^{ss}(\boldsymbol{k},\boldsymbol{k'})$. Additionally, the number density of an object and the weights assigned to it cannot depend its position. This means that $\bar{n}(\boldsymbol{r}) = \bar{n}$ and $w(\boldsymbol{r}) = w$ are constant. Hence,

$$G_{p,\ell}(0) = \int d^3r\bar{n}^p w^\ell = \bar{n}^p w^\ell V \tag{5.63}$$

where V is the volume of the region in which the covariance is being measured. Similarly

$$V_{\rm eff}(k_i) = \frac{\left(\int d^3r n^2 w^2\right)^2}{\int d^3\boldsymbol{r}\bar{n}^4w^4\left(1+\frac{1+\alpha}{\bar{n}(\boldsymbol{r})P(k_i)}\right)^2} = \frac{V}{\left(1+\frac{1}{\bar{n}P(k_i)}\right)^2}. \tag{5.64}$$

Substituting this pair of equations into the bin-averaged small-scale covariance from Eq. 5.59 results in the following solution for the covariance matrix without a window function,

$$\begin{aligned} C^{\rm no-win}(k_i,k_j) = \frac{\bar{T}(k_i,k_j)}{V} + \frac{2(2\pi)^3}{V_{k_i}V}\left(\bar{P}(k_i)+\frac{1}{\bar{n}}\right)^2\delta^D(k_i-k_j) \\ + \frac{1}{\bar{n}V}\left(4\bar{B}(k_i,k_j)+\bar{B}(0,k_j)+\bar{B}(k_i,0)\right) \\ + \frac{2}{\bar{n}^2V}\left(\bar{P}(k_i)+\bar{P}(k_j)+\bar{P}(k_i,k_j)\right)+\frac{(1+\alpha^3)}{\bar{n}^3V} \end{aligned} \tag{5.65}$$

This expression is the goal of this section, and is exactly that found by Meiksin & White (1999) and Scoccimarro et al. (1999). This expression has been used by many others investigating the covariance matrix, such as de Putter et al. (2012), Takada & Hu (2013) and Li et al. (2014a). It shows the expected volume scaling of the covariance matrix. Increasing the volume over which the power spectrum is measured directly decreases the error on the power spectrum, as there are more modes to average over to obtain the estimated power. The fact that the above expression arises as a result of the full expression for the covariance matrix in the absence of the window function helps validate Eq. 5.46.

The fidelity of Eq. 5.65 can be easily demonstrated. In the Gaussian regime with no shot-noise, only the term proportional to the power spectrum squared remains. For a set of Gaussian realizations, created using the method given at the beginning of Chapter 2, the measured covariance should match this analytic prediction. This is shown in Figure 5.1. In this figure the measured covariance (squared rooted to reduce the dynamical range) from a set of 500 Gaussian realizations with volumes $(640\,h^{-1}\,\mathrm{Mpc})^3$ and $(1280\,h^{-1}\,\mathrm{Mpc})^3$ and bins of width $\Delta k = 0.01\,h\,\mathrm{Mpc}^{-1}$ and $\Delta k = 0.04\,h\,\mathrm{Mpc}^{-1}$ is compared to the predictions from Eq. 5.65 in the Gaussian/no shot-noise regime. The realisations are based on an input power spectrum generated using CAMB and a galaxy bias of 2 is introduced to increase the signal-to-noise. The agreement between the two is exact within the limits of noise in the measured covariance matrix arising from using a finite number of realisations.

As can be seen from Fig. 5.1 increasing the volume and the bin width decreases the covariance. An increase in volume means that there are more independent modes to average over when estimating the power spectrum, hence the error is reduced. Similarly, when using wider bins, the number of independent modes that are available for averaging in each bin is increased.

Without shot-noise or non-Gaussianity, the covariance matrix as measured from a simulation is exactly diagonal. However the presence of trispectrum and higher order shot noise terms adds in an off-diagonal component. Additionally, the small scale diagonal covariance can be increased significantly by shot-noise (both first and higher order terms) and non-Gaussianity. These effects are demonstrated in Figures 5.2 and 5.3.

The first of these Figures shows the measured power spectrum error from a set of Gaussian Realizations calibrated against the power spectrum from the unmasked, unsubsampled MGS mock catalogues used elsewhere in this work and detailed in Chapter 3. This is compared to the diagonal covariance measured from the MGS simulations themselves. As the MGS simulations used here are neither masked nor subsampled, there is no window function, however the mocks used are the *galaxy* mocks (they have had the HOD applied) not the dark matter simulations. This means that there should be considerable

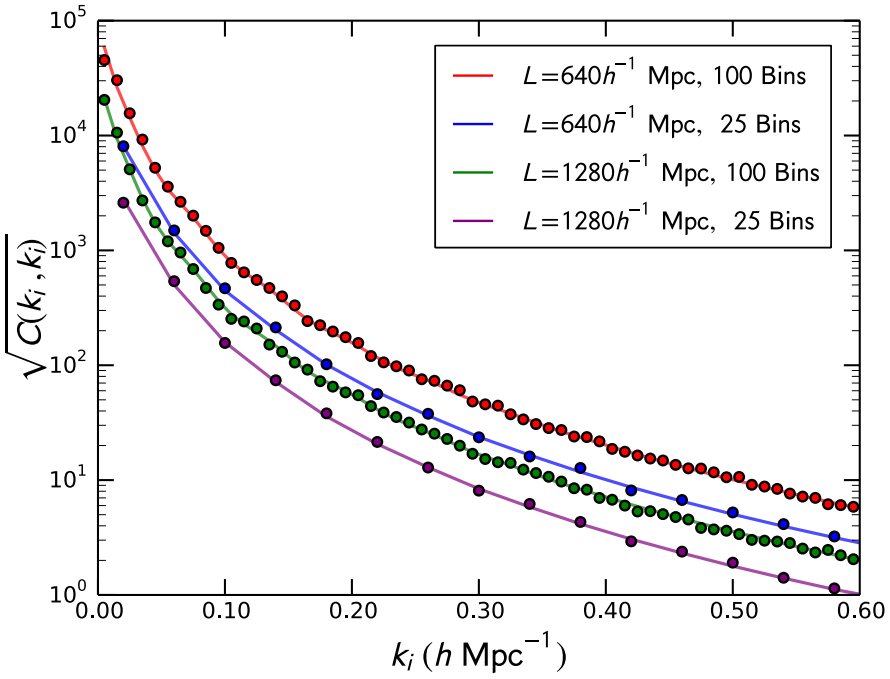


Figure 5.1: The error on the power spectrum from sets of 500 Gaussian realisations with different volumes and measurement bin widths. Points denote the measurements whilst the solid lines show the theoretical predictions. Increasing the bin width and the volume decreases the covariance as there are more modes in each bin to average over.

shot-noise and obvious non-Gaussian contributions to the power spectrum. In all these figures the error on the covariance matrix was estimated using bootstrap with random resampling.

In Figure 5.2 there is an obvious difference in the power spectrum errors between the Gaussian fields and non-linear mocks. Although both are in agreement on large scales, and with the theoretical Gaussian/no shot-noise predictions, the presence of shot-noise and non-Gaussian components significantly increases the covariance even on scales $k \approx 0.1\,h\,\mathrm{Mpc}^{-1}$. In order to distinguish these two additional contributions, the Gaussian prediction in the presence of 'first' order shot noise (i.e., still neglecting terms $\mathcal{O}(n^{-2})$ or higher) is also included as a dashed line. This shot noise contribution serves to change the effective volume such that there is a balance between the shot-noise and cosmic variance limited regimes. Whilst including this shot-noise does increase the theoretical covariance, the theoretical prediction still underestimates the covariance on scales $k > 0.1\,h\,\mathrm{Mpc}^{-1}$, indicating the measured covariance has significant non-Gaussian contributions that cannot be neglected.

To highlight this, Figure 5.2 also shows the ratio of the measured covariances against the Gaussian predictions with and without shot-noise. The ratio for the Gaussian Realisations divided by the Gaussian prediction is 1, as expected. The MGS covariance divided by the prediction without shot-noise differs by a factor of ≈ 10 on non-linear scales and the MGS covariance divided by the prediction including shot-noise still differs by a factor of 2. Hence the trispectrum, bispectrum and higher order shot-noise components serve to double the non-linear diagonal covariance and even if one neglects off-diagonal components of the covariance matrix these should still be included within the diagonal terms.

Figure 5.3 shows slices of the correlation matrix measured from the unmasked MGS simulations and the Gaussian realisations. The correlation matrix for the full, masked MGS simulations was previously plotted in Chapter 3. Slices from the correlation matrix for the unmasked mocks are plotted here to demonstrate that even in the absence of a mask, there is significant off-diagonal covariance in the power spectrum compared to that in the Gaussian realisations, caused by non-Gaussianity and higher-order shot noise. Neglecting this when calculating likelihoods or forecasts could result in significant underestimation of statistical errors.

There is one further effect to consider when measuring the covariance from a set of simulations. In recent literature it has been well documented that the presence of coupling between long (on the order of the simulation size) and small scale modes increases the covariance on small scales (Takada & Hu, 2013; Li et al., 2014a). Hence a simulation of a given volume will not return the 'true' covariance due to the absence of modes larger than

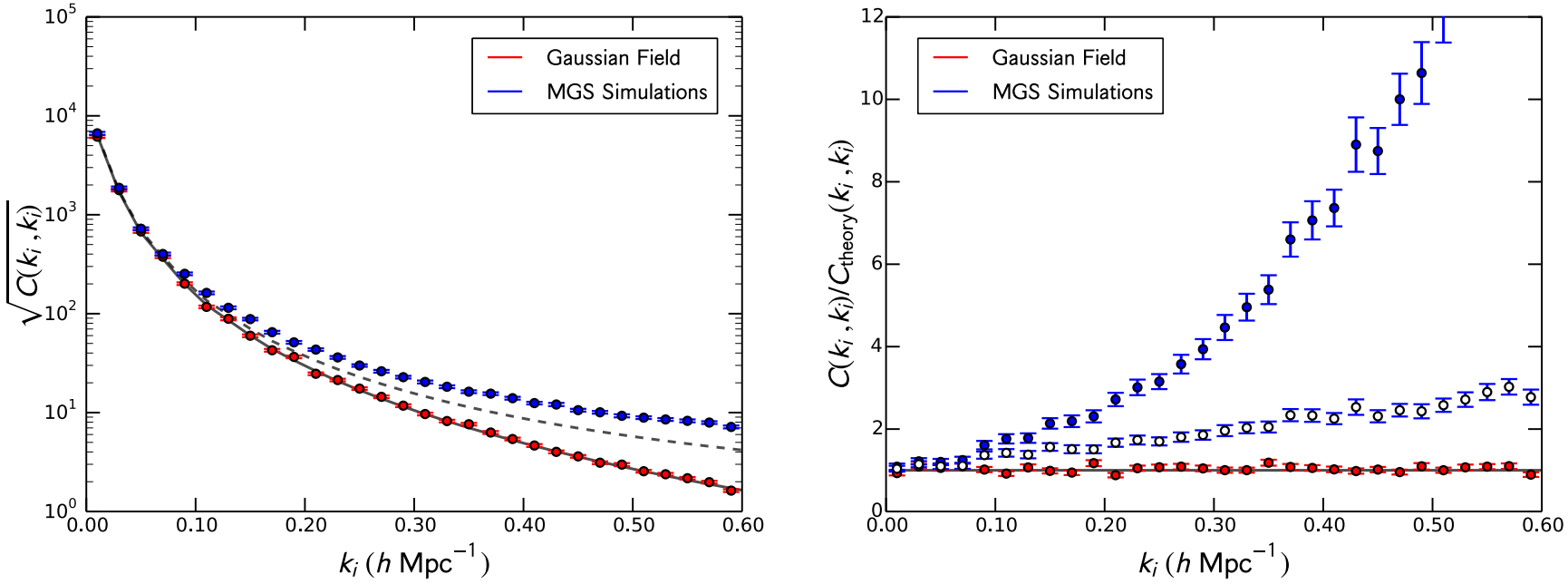


Figure 5.2: *Left*: The error on the power spectrum from 500 Gaussian Realisations (red) and from 500 of the unmasked, cubic galaxy mocks (blue) created for the analysis of the MGS detailed in Chapters 3 and 4. The black lines show the theoretical predictions for Gaussian fields with (dashed) and without (solid) the $\mathcal{O}(\bar{n}^{-1})$ shot noise correction. *Right*: The ratio between the measured covariance and the theoretical predictions. The solid points show the ratio using the prediction without shot-noise, whilst the open points include the dominant shot-noise component. The line gives the expected value of 1 for the Gaussian fields. Though all the measurements and predictions agree on large scales, the small-scale disparity between the Gaussian realisations, the analytic prediction including shot-noise, and the full non-linear simulations highlights the relative importances of shot-noise and the higher-order/non-Gaussian components of the covariance matrix.

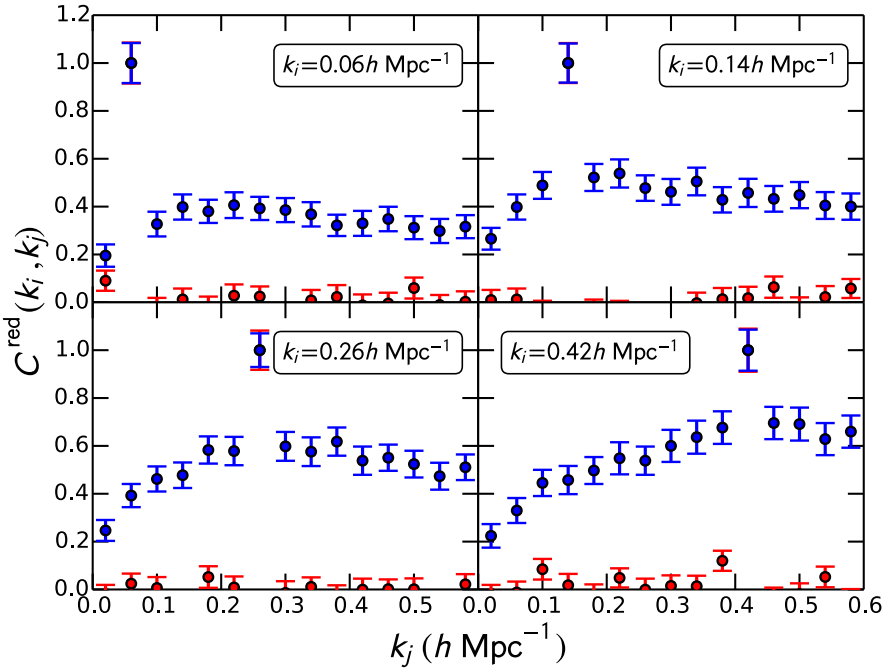


Figure 5.3: Slices of the correlation matrix of the power spectrum, defined as $C^{red}_{i,j} = C_{i,j}/\sqrt{C_{i,i}C_{j,j}}$, measured from the unmasked MGS simulated galaxy fields (blue) and Gaussian realisations (red). Slices covering a range of scales from the linear to non-linear are plotted, and in all there is a contribution to the off-diagonal covariance matrix from non-Gaussianity and higher order shot-noise.

the simulation box. In their work this phenomenon is called 'supersample' covariance, a name which will be adopted here. The next section explains this in more detail and also provides a way to correct for this which allows the volume scaling of the small simulation to return the correct, 'true' covariance matrix.

5.6 Supersample Covariance

The non-linear nature of gravitational evolution intimately couples long and short wavelength density fluctuations, introducing covariance between these modes. Due to the cosmological principle, on the very largest scales the density fluctuations should tend to zero. However, when observations of the universe are made, the finite size of the survey means that there may be density fluctutations larger than the survey which couple with modes inside the survey. Though these long wavelength perturbations cannot be measured directly, their interaction with the sub-survey modes still leaves additional information within the covariance matrix. This additional information is commonly known as beat-coupling, halo sample variance and, as will be adopted herein, supersample covariance.

The effect of super-survey modes on the power spectrum covariance matrix was originally studied by Hamilton et al. (2006) and Rimes & Hamilton (2006). Hu & Kravtsov (2003) also investigated the effect of these modes on the number counts of halos. Since then there have been many investigations into the nature of supersample covariance as well as its, possibly measurable, information content (Sefusatti et al., 2006; Takada & Bridle, 2007; Sato et al., 2009; Takada & Jain, 2009; Takahashi et al., 2009; de Putter et al., 2012; Kayo et al., 2013; Takada & Hu, 2013; Li et al., 2014a,b; Takahashi et al., 2014).

In particular, Takada & Hu (2013) give a detailed mathematical description of supersample covariance and its origin. In their work they find that supersample covariance arises from the convolution between the survey window and the trispectrum, given by the first term in eq. 5.46, in the squeezed limit. In this limit, the quadrilaterals that make up the trispectrum consist of two, nearly equal and opposite, long wavelength modes. As the other two orthogonal modes are small, the trispectrum in this regime acts as the modulation of two short wavelength power spectra by some background mode δ_b. In the peak-background split framework (Kaiser, 1984; Cole & Kaiser, 1989) this trispectrum term introduces a long wavelength density perturbation which modulates the amplitude of small scale pairs and changes the relative abundances of local peaks above the collapse threshold. Mathematically, the clustering quantity of interest is

$$T(\boldsymbol{k}, -\boldsymbol{k} + \boldsymbol{q}_{12}, \boldsymbol{k}', -\boldsymbol{k}' - \boldsymbol{q}_{12}) \approx T(\boldsymbol{k}, -\boldsymbol{k}, \boldsymbol{k}', -\boldsymbol{k}') + \frac{\partial P(k)}{\partial \delta_b} \frac{\partial P(k')}{\partial \delta_b} P^L(q_{12}) \quad (5.66)$$

As the mode q_{12} has a long wavelength, the power spectrum of this mode is the linear

power spectrum.

Substituting the above expression for the trispectrum into the covariance matrix including the window function, Eq 5.46, and then taking the small-scale limit, results in a modified expression for the small-scale covariance including the effects of modes larger than the survey, $C^{ssc}(k_i, k_j)$, based on the original small-scale covariance, $C^{ss}(k_i, k_j)$,

$$C^{ssc}(k_i, k_j) = C^{ss}(k_i, k_j) + (\sigma_b^L)^2 \frac{\partial P(k_i)}{\partial \delta_b} \frac{\partial P(k_j)}{\partial \delta_b} \tag{5.67}$$

where

$$(\sigma_b^L)^2 = \int \frac{d^3k}{(2\pi)^3} \frac{|G_{2,2}(\boldsymbol{k})|^2}{(G_{2,2}(0))^2} P^L(k), \tag{5.68}$$

is the variance of the background mode δ_b within the survey window. This is closely related to the well known quantity σ_8, but instead gives the linear matter variance within some survey, as opposed to a sphere of radius $8\mathrm{Mpc}$. As any reasonable survey volume will be much larger than the volume of such a sphere, the value of $(\sigma_b^L)^2$ will be considerably smaller. For a cubic region with side length L, the variance takes the form

$$(\sigma_b^L)^2 = 8 \int \frac{d^3k}{(2\pi)^3} P^L(k) \mathrm{sinc}^2\left(\frac{k_x L}{2}\right) \mathrm{sinc}^2\left(\frac{k_y L}{2}\right) \mathrm{sinc}^2\left(\frac{k_z L}{2}\right). \tag{5.69}$$

The formalism of Takada & Hu (2013) provides a useful way of characterising the effect of supersample covariance on cosmological measurements and of disentangling and utilising the signal from modes outside the survey in obtaining cosmological constraints (Li et al., 2014b). Of direct interest in this chapter however is the work of Li et al. (2014a) who investigate the effect of supersample covariance on *simulations*.

Unlike in surveys, where modes outside the volume encode information inside the volume, periodic simulations have no external modes. These are implicitly set to zero along with the average overdensity. Hence the covariance measured from an ensemble of simulations will be lower than that measured from an ensemble of real surveys of the same volume. Similarly the covariance of a set of small volume simulations will be lower than that of a set of larger simulations (after volume scaling) due to the absence of modes larger than the small volume. Some of these are present in the large volume simulation. However, on top of this the large volume simulation will be missing modes that would be present in an even larger simulation, though the effect of super-survey modes will diminish as larger and larger volumes are simulated.

Hence an estimate of the true covariance for some measured survey requires the inclusion of modes larger than the simulation volume. This is identified in Li et al. (2014a) who find that a set of small volume simulations can significantly underestimate the covariance even on moderately large ($k \approx 0.1$) scales. They also investigate analytic methods of including modes larger than the simulation volume. If the volume scaling method presented within this chapter is to work effectively, a method for introducing 'larger than box' modes into the small volume simulations will also have to be included here.

5.6.1 The Separate Universe Approach

Unlike the analytic method of Li et al. (2014a) this chapter presents a simple correction that can be made when running ensembles of simulations to correct for the absence of supersample covariance. This is based on the separate universe approach of Sirko (2005). The supersample covariance is described by the response of the power spectrum to a large scale background mode. However, rather than measuring this from a set of simulations with different background modes, the background mode can be included in the simulation parameters themselves, such that the measured covariance of the simulations already includes this response. This method has been presented several times in the past under different guises (Frenk et al., 1988; Tormen & Bertschinger, 1996; Cole, 1997) and is here presented in a way that is easy to implement and can be used with the L-PICOLA code presented in Chapter 2.

In the separate universe approach, the background mode is treated as a density contrast, which is then absorbed into the mean density of the simulation. Hence the mean matter density within a given simulation box with some background mode is related to the mean density of the ensemble via

$$\bar{\rho}_m^{\rm box} = (1+\delta_b)\bar{\rho}_m. \tag{5.70}$$

As shown in Chapter 1, Section 1.1.4, the matter density scales in proportion to the inverse-cube of the scale factor. Hence the scale factor of the simulation evolves as

$$a^{\rm box} = \frac{a}{(1+\delta_b)^{1/3}} \approx a\left(1-\frac{\delta_b}{3}\right) \tag{5.71}$$

compared to the ensemble. This change in the scale factor modifies the 'local' Hubble parameter of each simulation, which can be seen by taking the time derivative of Eq. 5.71,

$$\frac{\dot{a}_{\rm box}}{a_{\rm box}} = \frac{\dot{a}}{a} - \frac{\dot{\delta}_b/3}{1-\delta_b/3}. \tag{5.72}$$

To first order, the 'local' hubble parameter then becomes

$$H_{\rm box}^2 = H^2 - \frac{2}{3}H\dot{\delta}_b. \tag{5.73}$$

Fully incorporating the change in the scale factor and Hubble parameter into different simulations requires modifying the dark matter equation of state. However a simpler method is to use the Friedmann equation to calculate the 'local' density parameters for each simulation and then use these as input for each run in the ensemble. The Friedmann equation for each simulation is

$$H_{\rm box}^2 = H_{0,\rm box}^2\left(\frac{\Omega_{m,0,\rm box}}{a_{\rm box}^3} + \Omega_{\Lambda,0,\rm box} + \frac{\Omega_{k,0,\rm box}}{a_{\rm box}^2}\right). \tag{5.74}$$

However, at early times all the simulations within an ensemble should coincide, $a_{\rm box} = a$. In this regime, $\Omega_{m,{\rm box}} h^2_{\rm box} = \Omega_m h^2$, in which case a suitable parameterisation for the density parameters within each simulation box is

$$\begin{aligned} H_{0,{\rm box}} &= H_0(1+\phi)^{-1}, \\ \Omega_{m,0,{\rm box}} &= \Omega_{m,0}(1+\phi)^2, \\ \Omega_{\Lambda,0,{\rm box}} &= \Omega_{\Lambda,0}(1+\phi)^2, \\ \Omega_{k,0,{\rm box}} &= 1-(1+\phi)^2(\Omega_{m,0}+\Omega_{\Lambda,0}). \end{aligned} \tag{5.75}$$

Substituting this parameterisation into the local Friedmann equation for each simulation, expanding the terms and only keeping those that are first order in δ_b or ϕ gives

$$H^2_{\rm box} = H^2 + \frac{2H_0^2\phi}{a^2} + H_0^2\delta_b\left(\frac{\Omega_{m,0}}{a^3} + \frac{2}{3}\frac{\Omega_{k,0}}{a^2}\right). \tag{5.76}$$

Comparing this to Eq. 5.73 shows that the change in the density and Hubble parameters for each simulation can be written purely as a function of the background density and its time derivative.

The time derivative of the background mode is simple to calculate. As the background mode is expected to be linear it evolves based on the linear growth factor, $\delta_b(a) = D_1(a)\delta_b(0)$, and hence the time derivative of this is

$$\dot{\delta}_b = \frac{1}{\dot{a}}\frac{D_1'}{D_1}\delta_b. \tag{5.77}$$

Because the background mode scales as the growth factor, this is independent of the exact scale factor at which the growth factor and background modes are defined.

An expression for the differential of the growth factor with respect to the scale factor, D_1' has already been presented in Chapter 2. Hence the time differential of the background mode is

$$\dot{\delta}_b = \frac{\Omega_{m,0}H_0^2\delta_b}{2Ha^3}\left(\frac{5}{D_1} - \frac{2\Omega_{k,0}}{\Omega_{m,0}} - \frac{3}{a}\right). \tag{5.78}$$

Combining this with Eqs. 5.73, 5.76 and a little arithmetic, results in

$$\phi = \frac{5\Omega_{m,0}}{6}\frac{\delta_b}{D_1}, \tag{5.79}$$

which finally relates the background mode to the input parameters of a given simulation and allows one to run an ensemble of simulations that include modes larger than the simulation volume. This is the result found by Sirko (2005).

Overall, as the background mode is large scale, it is expected to be drawn from Gaussian distribution. Hence to include supersample covariance in a set of simulations, the following steps must be taken:

1. Calculate the variance in the background modes based on the input linear power spectrum *at the simulation redshift* and the box size.

2. For each simulation draw a background mode from a Gaussian distribution with zero mean and standard deviation given by σ_b^L.

3. Evaluate the new cosmology and output redshift for each simulation, based on the values of δ_b and ϕ.

4. Run the simulations as normal.

This procedure should work for any simulation code, although care must be taken to ensure that all parameters that depend on the background mode are modified correctly. When using L-PICOLA for this there are several other parameters that must be modified. Firstly, the redshift at which timestepping begins is modified in the same way as the output redshift, to ensure that the amount of time spent evolving the dark matter field is correctly modified by the background mode. Secondly, the value of σ_8 that is passed to L-PICOLA must also be modified. The reason for this is *not* physical however, as the change in the growth of structure in each simulation has already been captured by the modifications to the input cosmology and output redshift. Rather this is due to the fact that L-PICOLA requires an input power spectrum at redshift zero and an associated value of σ_8 at redshift zero to generate the initial conditions. The power spectrum at the redshift of the initial conditions is then scaled back by the growth factor. As the cosmology of each simulation is slightly different, so too is the growth factor, such that the power spectrum amplitude at the redshift of the initial conditions, as calculated by L-PICOLA is *not* the same for each simulation.

As used in the separate universe derivation above, one would expect that the simulations should be coincident at high redshift. To then ensure that this *is* true, it is necessary to scale the value of σ_8 that is given to each L-PICOLA run. If L-PICOLA used an input power spectrum at some suitably high redshift, then this correction would not be necessary. The exact form of the modification to σ_8 is just the ratio of the normalised growth factors in the fiducial and 'box' cosmologies, i.e.,

$$\sigma_{8,\mathrm{box}} = \sigma_8 \frac{D_1^2(z_{sync}, \Omega_m, \Omega_\Lambda)}{D_1^2(0, \Omega_m, \Omega_\Lambda)} \frac{D_1^2(0, \Omega_{m,\mathrm{box}}, \Omega_{\Lambda,\mathrm{box}})}{D_1^2(z_{sync}, \Omega_{m,\mathrm{box}}, \Omega_{\Lambda,\mathrm{box}})} \tag{5.80}$$

Hence, for every L-PICOLA simulation, the value of σ_8 at z_{sync} matches.

5.6.2 Tests on L-PICOLA Simulations

The remainder of this section will be dedicated to showing that this procedure recovers the supersample covariance for an ensemble of L-PICOLA simulations. This section uses four sets of 500 cubic dark matter simulations with side lengths/number of particles $L = 512\,h^{-1}\,\mathrm{Mpc}, N = 256^3$ and $1024\,h^{-1}\,\mathrm{Mpc}, N = 512^3$, and with and without the correction for supersample covariance included. In all other aspects the simulations are

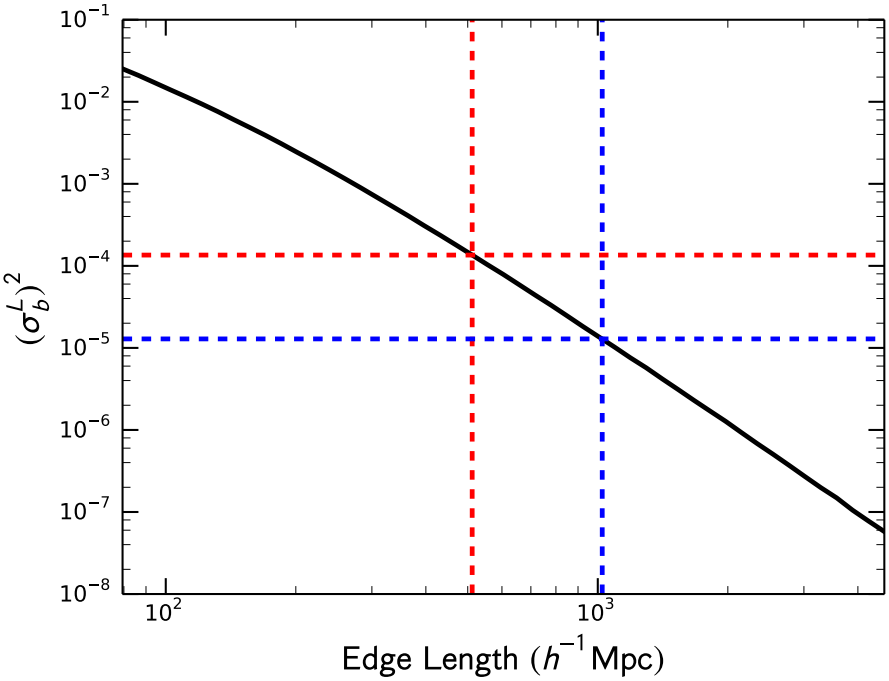


Figure 5.4: The variance in the background modes $(\sigma_b^L)^2$ for cubic simulations with differing volumes. The red and blue lines correspond to the $L = 512\,h^{-1}\,\mathrm{Mpc}$ and $L = 1024\,h^{-1}\,\mathrm{Mpc}$ simulation used in Section 5.6.2 to test the supersample covariance correction.

identical. Only dark matter simulations are used to test this correction as the non-linear nature of the supersample covariance means that it is largely hidden by shot-noise in a galaxy mock catalogue. The simulations are generated using a linear power spectrum from CAMB and a flat fiducial cosmology with $\Omega_m = 0.31$, $n_s = 0.96$ and $\sigma_8 = 0.83$. They are evolved using the modified COLA timestepping method with 11 timesteps from an initial redshift of z=9.0 up to z=0.0.

The first step is the calculation of $(\sigma_b^L)^2$. Figure 5.4 shows the value of the $(\sigma_b^L)^2$ for cubic simulations of differing sizes. The values for the two simulation sizes used in this section are marked. This figure shows that the variance rapidly decreases as the simulation gets larger, but because of its logarithmic nature it can have a large effect on the covariance matrix recovered even from two sets of simulations that do not differ drastically in size.

Using the values of $(\sigma_b^L)^2$, values for δ_b are drawn from a Gaussian distribution for each of the 1000 simulations that have the supersample covariance correction, 500 large and 500 small. These values of δ_b are then used to modify the input parameters for each simulation. The uncorrected simulations are run using only the fiducial cosmology. The power spectrum for each simulation is calculated using a number of cells equal to the

length, i.e., a constant cellsize of $1\,h^{-1}\,\mathrm{Mpc}$, to ensure fair comparison between the small and large volume simulations. The power spectra and covariance matrices are calculated using 25 bins in the range $0.0 < k < 2.0$. The errors on the covariance matrix are calculated using bootstrap resampling.

Figure 5.5 shows the result of the supersample covariance correction. All three panels show the ratio of the small and large uncorrected covariance, and the small, corrected covariance, against the large, corrected covariance. One should expect that both the small and large volume simulations containing the correction converge on small scales to the true covariance, and that both the small and large uncorrected simulations are missing small scale covariance due to the absence of supersample modes. The large uncorrected box should be closer to the true covariance than the small uncorrected box.

These expected results are exactly that shown in Figure 5.5. Although the errors are quite large, the absence of supersample modes in the uncorrected large and small boxes is very noticeable. Additional small scale covariance is added by the separate universe approach, and though the true covariance is, strictly speaking, unknown, both the large and small corrected boxes converge to the same covariance, even though the magnitude of the correction applied to these two sets is very different.

5.7 Combining Analytic Estimates of the Covariance Matrix with Simulations

Throughout the previous sections, even in a periodic simulation, the importance of the non-Gaussian, higher-order shot noise and super-sample components has been demonstrated. These terms are often neglected in parameter estimation, likelihood fitting of the power spectrum and Fisher matrix forecasts (Tegmark, 1997), and though the low redshift and number density of the MGS will exacerbate these effects due to the significant non-linear evolution of structure, it will certainly be necessary to include these in the analysis of next generation surveys. This presents a problem however, as the bispectrum and trispectrum (and to some extent the power spectrum) are extraordinarily difficult to model analytically on non-linear scales and in redshift space. The most accurate way to do this, and hence to calculate the covariance matrix on these scales, is still to use simulations.

In the face of these difficulties, it is beneficial to use a combination of the theoretical model described in this chapter and numerical simulations to estimate the covariance. This solves the problem of modelling the small scales, whilst allowing the large scales to be evaluated analytically, without the need for large volume simulations. These two approaches can be combined using knowledge of the behaviour of the covariance matrix when using different simulation volumes and bin sizes.

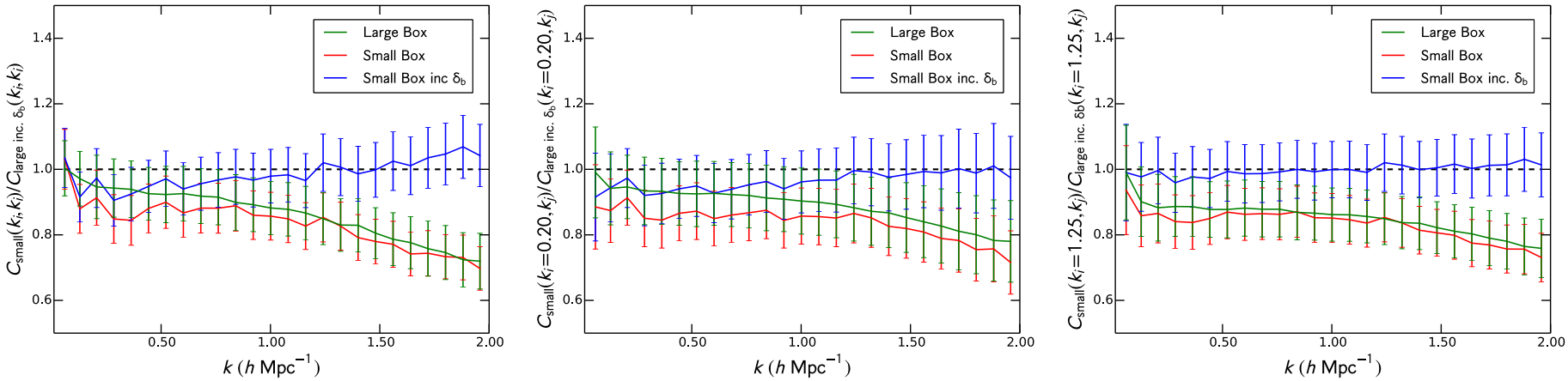


Figure 5.5: The ratio of the covariance matrices measured from the uncorrected large and small sets of simulations and the corrected small scale simulations against the corrected large scale simulations, detailed in Section 5.6.2. The left panel shows the diagonal elements, whilst the middle and right panels show slices through the covariance matrices at $k = 0.2\,h\,\mathrm{Mpc}^{-1}$ and $k = 1.25\,h\,\mathrm{Mpc}^{-1}$ respectively. The fact that the large corrected and small corrected boxes return the same covariance indicates that the absence of supersample modes in the uncorrected simulations is being fixed by the separate universe approach detailed in this chapter.

This section shows how the covariance matrix measured from a set of small volume simulations can be rescaled based on the analytic expectations for the behaviour of the covariance matrix. This means that the computational requirements to reach a given covariance matrix precision can be greatly reduced. Unlike other methods for achieving this however, no fitting functions are required and there are no additional parameters to calibrate or rely on. The same well-understood techniques that are currently used to generate a set of mock catalogues, such as those for halo-finding and populating the halos with galaxies, can still be used. All that is required is a rescaling of the measured covariance.

5.7.1 Cubic Simulations

In the absence of a window function it has been shown how the covariance matrix changes with simulation volume. Hence it is possible to simulate smaller volumes and then scale the covariance matrix based on the simulation volume used and the simulation volume which would have ideally been used. To show this two sets of mock catalogues were run. The first is simply the MGS mock catalogues from Chapters 3 and 4 which have been used previously in this chapter. The second set is 1000 galaxy simulations run using the same cosmology, initial power spectrum and HOD model, but with a volume and number of dark matter particles one-eighth that used for the MGS simulations. Hence the simulation specifications have been reduced from box length $L = 1280\,h^{-1}\,\mathrm{Mpc}$ and number of particles $N = 1536^3$ to $L = 640\,h^{-1}\,\mathrm{Mpc}$ and $N = 768^3$.

Figure 5.6 compares the power spectra and their errors from the two sets of simulations. As would be expected the power spectra of the two sets agree exactly within the error bars. The diagonal elements of the covariance matrix do not agree with each other, however they both agree to approximately the same degree with their respective analytic predictions, with and without shot-noise. Hence, this would lead one to suggest that the ratio between the analytic predictions would be the same as the ratio between the measured covariance matrices, and matching the larger volume covariance matrix is as simple as scaling the smaller volume covariance by the analytic volume ratio.

The accuracy of this is demonstrated in Fig. 5.7 where the ratio of the measured covariances and the analytic predictions are shown for both the diagonal and off-diagonal elements of the covariance matrix. The agreement between the two is within the error bars for all elements plotted and on all scales up to $k \approx 0.3\,h\,\mathrm{Mpc}^{-1}$. After this the ratio between the covariance matrices has a tendency to be larger than the analytic prediction. This is most obvious for the diagonal elements where the error on the covariance is smaller. The small volume simulation has less covariance than the larger simulation when the volume scaling is taken into account. This is due to the absence of super-

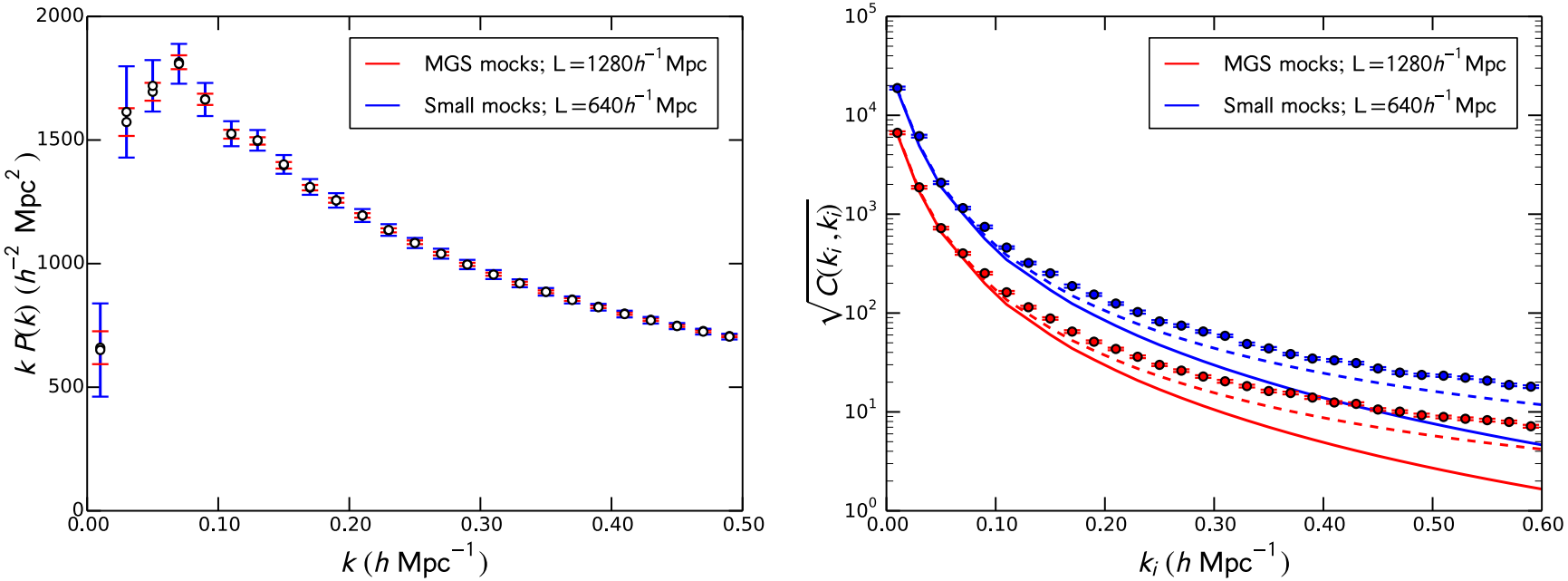


Figure 5.6: *Left*: The power spectrum from 500 of the unmasked, cubic galaxy mocks created for the analysis of the MGS (red) and from 500 simulations which use one-eighth the volume and number of particles, but are otherwise identical. The power spectra agree within the error bars as would be expected. *Right* The measured covariance and the theoretical predictions for the two sets of simulations. The solid lines give the analytic covariance without shot noise, whilst the dashed line is the prediction when the first-order shot-noise contribution is included via the *effective* volume.

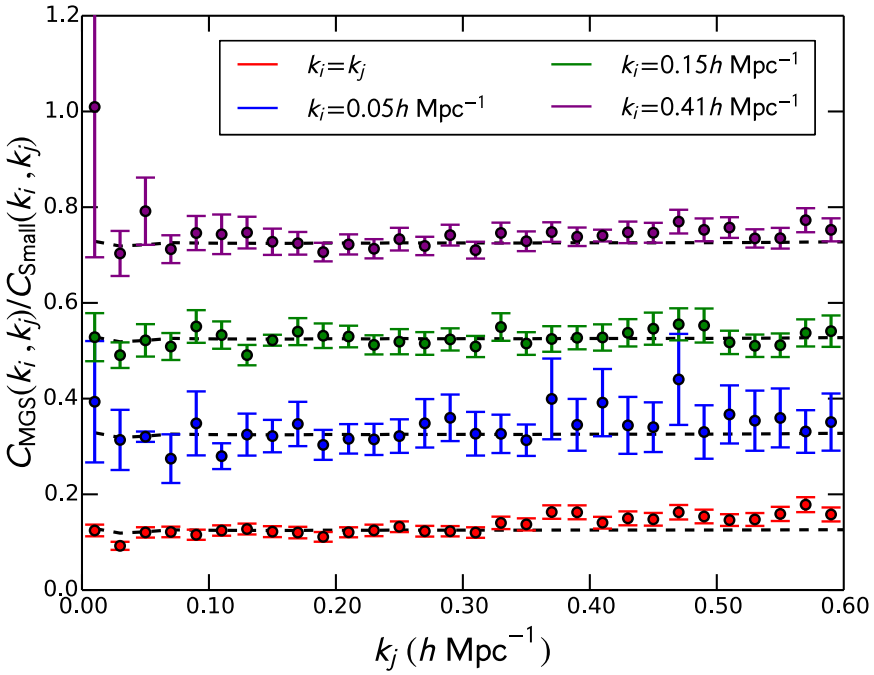


Figure 5.7: The ratio between the covariance measured from two sets of MGS-based galaxy mocks that differ only in the number of particles and box volume. The large simulations are eight times larger and contain eight times more particles (and hence approximately eight times more galaxies). Different colors correspond to different slices of the covariance matrices, and each ratio has been displaced by 0.2 compared to the previous to aid visualisation. Dashed lines show the expected ratio based on the analytic predictions (similarly displaced). The agreement between the simulations and theory is remarkable, except on non-linear scales where supersample covariance becomes important. This is detailed more in Section 5.6

sample modes. A correction for this has been given above, but has not been included here to highlight the effect of neglecting these on the volume scaling procedure. The modification to the input parameters to include supersample modes will be included as part of the next section.

5.7.2 Masked Simulations Including Supersample Covariance

The volume dependence of the covariance matrix provides a simple way of rescaling a set of cubic simulations. However it can also be shown that this works on small scales for a set of masked mock catalogues., i.e., one can rescale a set of small cubic galaxy mocks to match the covariance that would be obtained from a set of masked mocks with redshift dependent number density. At first this may not seem obvious due to the differences in number density between the masked and periodic simulations. Even on small scales

where there is no convolution with the window function, the position-dependent number density and weights applied to the mocks still affect the covariance. In a set of small volume cubic simulations these position-dependent effects cannot be fully included as the survey will not fit inside the simulated volume.

However, there is a way to include these effects analytically. The small scale masked and simulation covariances measured in some k-bins are given by Eqs. 5.59 and 5.65. If one assumes that the small scale power spectrum, bispectrum and trispectrum should be identical between the masked and small volume, cubic mocks then these equations can be combined. Treating the diagonal and off-diagonal terms separately this results in the relatively simple expression for the small-scale masked covariance,

$$C^{ss}(k_i,k_j) = \begin{cases} \frac{V_{eff}^{\rm no-win}(k_i)}{V_{eff}^{\rm ss}(k_i)} C^{\rm no-win}(k_i,k_j) + C^{\rm res}(k_i,k_j) & i=j \\ \frac{(G_{2,2}^{\rm no-win}(0))^2}{(G_{2,2}^{ss}(0))^2} \frac{G_{4,4}^{ss}(0)}{G_{4,4}^{\rm no-win}(0)} C^{\rm no-win}(k_i,k_j) + C^{\rm res}(k_i,k_j) & i\neq j \end{cases} \tag{5.81}$$

The above equation shows that, as long as the residual $C^{\rm res}(k_i,k_i)$ terms are small, the small scale covariance of a set of masked galaxy mocks is simply some rescaling of the covariance matrix measured from the small-volume cubic simulations. The diagonal terms scale as the ratio of the effective volumes, whilst the off-diagonal terms scale simply as some scale-independent volume ratio that does not include the trade off between cosmic variance and shot noise.

The residual component of the covariance matrix details the difference between the number densities of the two ensembles of mocks and the fact that the cubic simulations do not include any position-dependence. For a volume limited survey with constant number density this term would be expected to reduce to zero as the number of the cubic simulations can be matched to the survey number density exactly. In fact even for a non-volume limited survey this term is expected to be close to 0 and can be taken as such. In truth, treating the diagonal and off diagonal parts separately, it takes the form

$$\begin{aligned} C^{\rm res,diag}(k_i,k_i) = \bar{T}(k_i,k_j) &\left(\frac{G_{4,4}^{ss}(0)}{(G_{2,2}^{ss}(0))^2} - \frac{V_{eff}^{\rm no-win}(k_i)}{V_{eff}^{\rm ss}(k_i)} \frac{G_{4,4}^{\rm no-win}(0)}{(G_{2,2}^{\rm no-win}(0))^2} \right) \\ &+ \left(4\bar{B}(k_i,k_j) + 2\bar{B}(0,k_j) \right) \left(\frac{G_{3,4}^{ss}(0)}{(G_{2,2}^{ss}(0))^2} - \frac{V_{eff}^{\rm no-win}(k_i)}{V_{eff}^{\rm ss}(k_i)} \frac{G_{3,4}^{\rm no-win}(0)}{(G_{2,2}^{\rm no-win}(0))^2} \right) \\ &+ \left(4\bar{P}(k_i) + 2\bar{P}(k_i,k_i) \right) \left(\frac{G_{2,4}^{ss}(0)}{(G_{2,2}^{ss}(0))^2} - \frac{V_{eff}^{\rm no-win}(k_i)}{V_{eff}^{\rm ss}(k_i)} \frac{G_{2,4}^{\rm no-win}(0)}{(G_{2,2}^{\rm no-win}(0))^2} \right) \\ &+ \left(1+\alpha^3 \right) \left(\frac{G_{1,4}^{ss}(0)}{(G_{2,2}^{ss}(0))^2} - \frac{V_{eff}^{\rm no-win}(k_i)}{V_{eff}^{\rm ss}(k_i)} \frac{G_{1,4}^{\rm no-win}(0)}{(G_{2,2}^{\rm no-win}(0))^2} \right), \end{aligned} \tag{5.82}$$

$$
\begin{aligned}
C^{\mathrm{res,off-diag}}(k_i,k_j) = &\Bigg(4\bar{B}(k_i,k_j)+\bar{B}(0,k_i)+\bar{B}(0,k_j)\Bigg) \\
&\times\left(\frac{G^{ss}_{3,4}(0)}{(G^{ss}_{2,2}(0))^2}-\frac{G^{\mathrm{no-win}}_{3,4}(0)}{(G^{\mathrm{no-win}}_{2,2}(0))^2}\frac{G^{\mathrm{ss}}_{4,4}(0)}{G^{\mathrm{no-win}}_{4,4}(0)}\right) \\
&+2\Bigg(\bar{P}(k_i)+\bar{P}(k_j)+\bar{P}(k_i,k_j)\Bigg) \\
&\times\left(\frac{G^{ss}_{2,4}(0)}{(G^{ss}_{2,2}(0))^2}-\frac{G^{\mathrm{no-win}}_{2,4}(0)}{(G^{\mathrm{no-win}}_{2,2}(0))^2}\frac{G^{\mathrm{ss}}_{4,4}(0)}{G^{\mathrm{no-win}}_{4,4}(0)}\right) \\
&+\Big(1+\alpha^3\Big)\left(\frac{G^{ss}_{1,4}(0)}{(G^{ss}_{2,2}(0))^2}-\frac{G^{\mathrm{no-win}}_{1,4}(0)}{(G^{\mathrm{no-win}}_{2,2}(0))^2}\frac{G^{\mathrm{ss}}_{4,4}(0)}{G^{\mathrm{no-win}}_{4,4}(0)}\right)
\end{aligned}
\tag{5.83}
$$

Using the expressions for the effective volume and $G_{p,l}$ terms, it is simple to show that this does indeed reduce to zero in the limit of constant $\bar{n}$ and w. In the scaling of the off-diagonal covariance, the trispectrum has been prioritised, such that the scaling perfectly accounts for the trispectrum term and there are no residual components. However based on the amount of shot-noise it may be more beneficial in some instances to prioritise the bispectrum terms instead.

As a proof-of-concept for this method, the masked, subsampled MGS mocks from Chapters 3 and 4 are used. The small mocks that will be scaled are the same as those used previously in Section 5.7.1, however now the correction for the missing super-sample covariance has been included in the same way as Section 5.6.2. 1000 of these small mocks are used to ensure precise measurements of the covariance matrix. When including the supersample covariance in these small volume *galaxy* mocks, the HOD parameters remain the same as we assume that the fundamental way galaxies populate halos of a given mass is 'universal' and does not depend on the local environment, however the mass of the particles in each realisation does change due to the slight change in the simulation parameters. Hence the mass of the halos in each realisation is slightly modified by the background modes. In practice, it is found that whether or not this effect is included has no noticeable effect on the covariance, however, as the correction to the halo mass is extremely small.

Figure 5.8 shows the result of applying Eq. 5.81 to these small volume mocks. Plotted in this figure is the ratio of the diagonal elements of the small volume and masked covariance matrices before and after applying the rescaling method. On scales $k > 0.15\,h\,\mathrm{Mpc}^{-1}$, the amplitude of the rescaled covariance matrix matches that of the masked simulations extremely well. Achieving good agreement on these scales highlights the power of this method, as modelling the redshift space trispectrum and bispectrum on such scales would be extremely difficult, however the use of simulations bypasses this problem with ease.

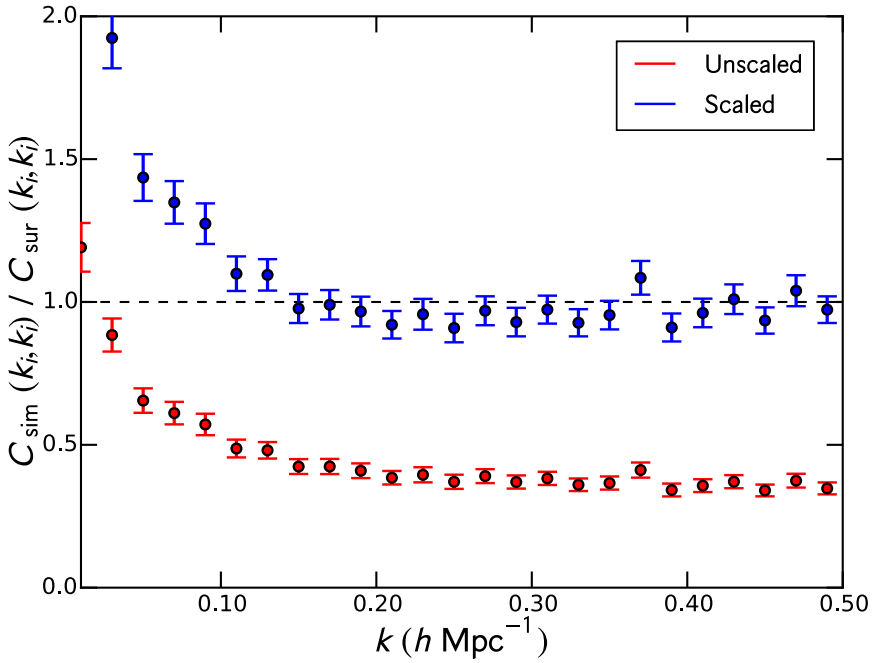


Figure 5.8: The ratio of the diagonal small-volume simulation covariance before and after analytic rescaling, compared to the diagonal covariance measured from the full set of masked, subsampled MGS mock catalogues. The small volume simulations have had the supersample covariance correction applied. There is very good agreement on small scales between the scaled small volume covariance and the 'true' covariance from the MGS mocks, however the small volume periodic simulations can be run much faster. The large scale deviation is due to the absence of the window function convolution in the small volume simulations which would otherwise reduce the large scale covariance.

However, on large scales the agreement is not so good and the scaled simulation covariance matrix overestimates the masked covariance. This is due to the window function, the convolution with which reduces power and diagonal covariance on large scales and introduces large off-diagonal covariance. The effect of the survey window is further shown in Figure 5.9, which shows slices of the correlation matrix for the masked MGS covariance matrix and the small volume covariance matrix before and after scaling. The masked MGS simulations have significantly more off-diagonal covariance that the small-volume simulations, which arises from the convolution with the window function. Hence to fully recover the covariance matrix of the masked simulations, the small volume simulations must be analytically convolved with the survey window function. The formalism detailed in this chapter provides a background upon which to base this analytic convolution, however a practical application of this convolution is not attempted here and is left as future work.

5.8 Summary and Application to Future Surveys

This chapter has presented a new method for estimating the covariance matrix from a set of simulations, which uses the information contained therein in a more efficient way. Rather than using the standard method of running a ensemble of large volume simulations that fit the full survey volume, this chapter advocates the use of a set of small volume cubic simulations to estimate the covariance matrix, which can then be rescaled simply through the use of an analytic formula.

This method will be of great use in estimating the power spectrum covariance matrices for future surveys such as Euclid (Laureijs et al., 2011), DESI (Levi et al., 2013), LSST (Ivezic et al., 2008) and SKA (Maartens et al., 2015), which will cover larger cosmological volumes than any survey to date. Producing 1000's of simulations which are detailed yet large enough for these surveys would prove immensely challenging, but the technique presented in this chapter provides an extremely promising alternative.

Section 5.1, expands on the motivation for this chapter. This section also introduces alternative methods for reducing the number of mock catalogues required to reach a given covariance matrix precision. Unlike these other methods however, the new technique in this chapter does not require or rely on any additional free parameters which must be fit or calibrated, and uses only well-understood simulation methods that have been applied to previous large structure surveys.

For instance, as just one example, for next generation surveys a single N-Body simulation could be run that fully captures the survey window and can be used to investigate the galaxy properties and fit a HOD model. Then an ensemble of much smaller volume simulations could be run and turned in cubic galaxy mocks using the same HOD

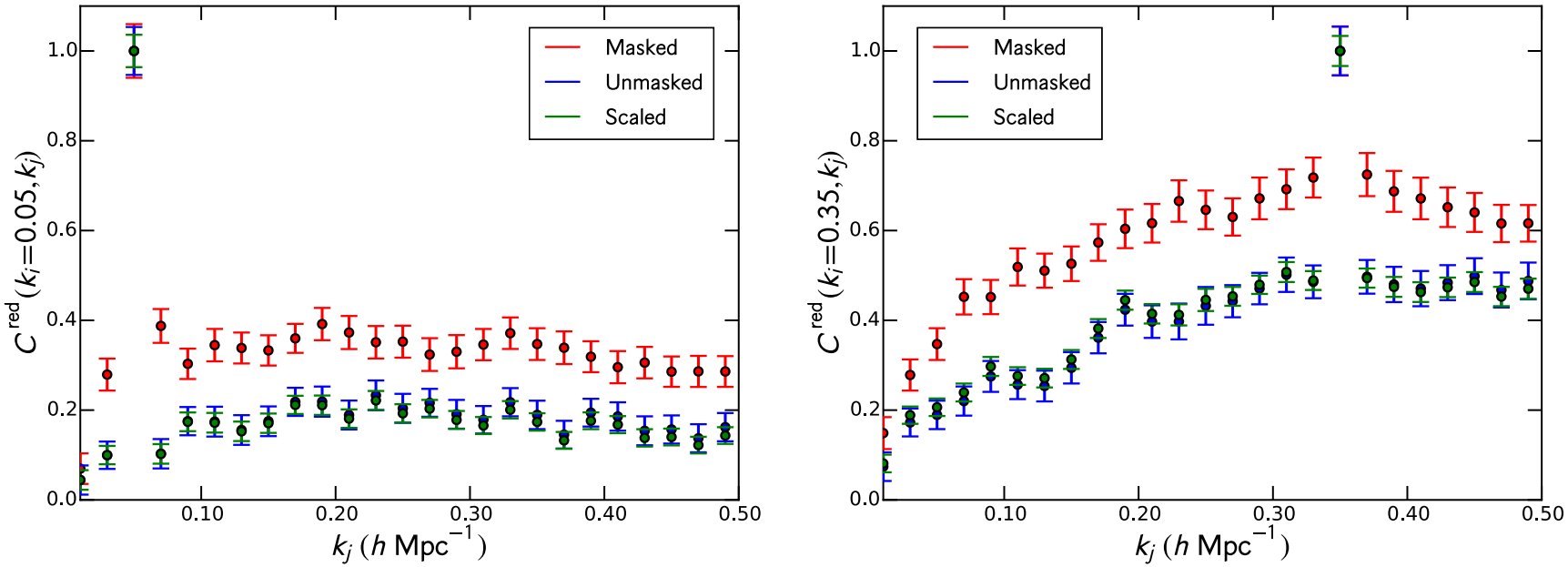


Figure 5.9: The correlation matrix for the masked MGS mocks and the small volume simulations before and after rescaling. The small volume simulations lack off-diagonal covariance compared to the masked simulations. The extra covariance is caused by the convolution with the window. The scaled simulations also lack this off-diagonal covariance, by the same relative amount. Increasing the off-diagonal covariance correctly requires analytic convolution of the covariance matrix with the survey window function.

model. The measured covariance from those small volume simulations is then simply scaled based on the properties of the full N-Body simulation and the small cubic simulations. This simple procedure is very similar to the standard method of covariance matrix estimation, except without the need to run many simulations the same size as the N-Body simulation. In fact, although the approximate code L-PICOLA has been used in this Chapter, if the boxes are made small enough one could even run the ensemble using a fully non-linear iterative N-Body code to recover the correct small scale information even on very non-linear scales.

To provide a theoretical basis for this technique Sections 5.2 and 5.3 show derivations of the power spectra and covariance matrix under the FKP formalism. The resultant covariance matrix formula includes all possible terms arising from Poisson shot-noise and non-Gaussianity, and describes the effect of the survey window function. To validate this expression the covariance matrix in the small-scale limit and in the absence of any window function is then derived and compared to existing studies in Sections 5.4 and 5.5.

One problem with using small volume simulations is the absence of information outside the simulation box, which can affect the small scale covariance and would be present in a larger simulation. This 'supersample' covariance was introduced in Section 5.6 where a correction for this was also shown. By modifying the simulation parameters of each run in an ensemble based on a Gaussian distributed background mode, the supersample covariance is recovered as shown by comparisons between large and scale simulations with and without the background mode correction.

Finally the separate theoretical equations and simulation techniques within this chapter are combined in Section 5.7. As a proof-of-concept this section shows that the covariance matrix from the masked, subsampled MGS mock galaxy catalogues detailed in Chapter 3 and 4 can be recovered extremely well on small scales using a set of cubic simulations which include the supersample covariance correction and that are 8 times smaller than those used to generate the MGS mocks. However some work still needs to be done on reproducing the large scale covariance, which is affected by the convolution with the survey window function. This is left as work for future studies.

Chapter 6

Conclusions

Over the three years in which the work in this thesis has been carried out, our understanding of the universe has increased dramatically, though in some aspects it remains the same as it did two decades ago. Recent high redshift CMB results from the Planck mission, and lower redshift LSS results from the BOSS have provided measurements of the cosmology of the universe at unparalleled precision. However, in all cases, the measurements still point towards a consensus ΛCDM cosmological model; where dark energy is described completely by a cosmological constant, the amount of primordial non-Gaussianity in the universe is consistent with zero and there exist only 3 flavours of neutrinos. In that sense the state of cosmology is much as it has been for the last 15 years.

The work presented within this thesis has shown new techniques that can be used to obtain cosmological results from LSS surveys and in being applied to existing datasets, has contributed to the status of the current consensus model. Chapter 2 presented a new code for the fast generation of dark matter simulations for the estimation of covariance matrices; Chapter 3 showed how this code can be combined with other techniques to produce mock galaxy catalogues for use in analysing data from the Sloan Digital Sky Survey; and Chapter 4 then showed the results of this analysis and the subsequent cosmological constraints. Although the work in these Chapters stands alone, combined, they detail another measurement that further validates the ΛCDM model.

That said, measurements of the dark energy equation of state, amplitude of non-Gaussian initial fluctuations and neutrino properties are still in their infancy. There exist a huge number of dark energy and inflationary models that are yet to be ruled out and as such the true nature of the universe is still very much undetermined.

Over the next decade missions such as Euclid (Laureijs et al., 2011), DESI (Levi et al., 2013), LSST (Ivezic et al., 2008) and SKA (Maartens et al., 2015) promise to measure the large scale structure of the universe in a variety of wavelengths and over huge volumes. The measurements these surveys will obtain will allow for extremely accurate measure-

ments of the dark energy equation of state across the whole of the dark energy dominated universe, measurements that could completely rule out ΛCDM or firmly cement it as the consensus cosmological model. As an example, Laureijs et al. (2011) forecast an improvement of a factor 10 on the precision of the time-independent dark energy equation of state, and a factor 30 improvement on the growth index from the Euclid survey. On top of this these surveys will allow probes of LSS to become the primary source of information on the amount of primordial non-Gaussianity and the sum of neutrino masses. For instance, in combination with CMB measurements the Euclid survey could produce a factor 50 improvement on current measurements of f_{NL} and factor 30 improvement on the sum of neutrino masses.

However, analysing the data from these surveys will require huge amounts of time and effort. The processing of the data alone will present a significant challenge just in terms of sheer data-volume, and in many cases existing computational and analysis techniques will not be good enough. The work presented in this thesis, though used mostly in the analysis of existing datasets, has been performed with these future surveys in mind. Features in the new code L-PICOLA such as the ability to include primordial non-Gaussianity and run lightcone simulations, presented and tested in Chapter 2, have been implemented to aid in covariance matrix estimation for next generation surveys. These features, implemented in a fast, accurate code, will be of utmost importance in understanding the statistical and systematic errors inherent in next generation surveys. Similarly, the work in Chapter 5 shows a new method for estimating the power spectrum covariance matrix in a way that solves one of the biggest computational problems facing next generation surveys, namely how to reconcile the need for high resolution simulations with the large volumes these surveys cover. With a little further work, the methods presented in these two chapters could become the most viable option for covariance matrix estimation for future surveys.

6.1 Future Work

Although much of the work in this thesis has been performed with next generation surveys in mind, there is still much that can be done to improve this and ensure that it stays at the forefront of those techniques that will be used.

6.1.1 Improvements to L-PICOLA

To begin, there are several improvements that could be made to L-PICOLA in the future. In terms of parallelisation, splitting the mesh into 'blocks' rather than 'slices' could improve both the speed and scalability of the code to large numbers of processors, however the need for additional MPI communication during the Fast Fourier Transforms means that

the level of improvement is indeterminate at this time. Furthermore one could imagine hybridising the code, using Open-MP and MPI such that communication between 'local' processors does not rely on slower MPI communication.

In terms of the physics behind L-PICOLA it would be simple to add in support for warm dark matter. Another obvious addition to the code would be to implement the spatial extension of the COLA method, presented by Tassev et al. (2015). Such an improvement would allow for the simulation of a large cosmological volume whilst only spending computational time evaluating the non-linear displacements for a small portion of that volume. Lightcone simulations within L-PICOLA in particular would greatly benefit from this as one would be able to simulate a small 'pencil-beam' region of the full lightcone and scale this up to the required simulation volume. Also, as this extended COLA method still requires the 2LPT displacements to be calculated for all the particles within the full volume, implementing this into the distributed-memory code would allow much larger cosmological volumes and higher particle densities to be simulated than the current shared-memory implementation of Tassev et al. (2015).

Additional small scale accuracy could be achieved by a suitable scaling of the mesh during the simulation, such as using a finer mesh at late times when the particles become more clustered. This would be particularly easy to implement as, in the optimal memory case, the mesh is deallocated and reallocated each time step anyway. Using an adaptive mesh for high density portions of the simulation, or the Tree-PM algorithm instead of the PM algorithm, could also be implemented though these methods would come at a cost to speed.

As L-PICOLA is so fast, for current applications the total CPU time taken to produce a mock galaxy catalogue is dominated by outputting and post-processing of (mainly reading in) the dark matter field, especially the creation of dark matter halos. This is exacerbated for lightcone simulations with replication as the simulation is effectively being output multiple times, resulting in large increases to the amount of time taken to output and process the data. This could be vastly improved by adding in a halo finder into L-PICOLA, either by identifying shell-crossing as it occurs during the simulation, or via the FoF algorithm. This would mean that the amount of time taken to output the data, and read it in for post-processing, could be reduced drastically.

6.1.2 Future work on mock catalogue production and RSD measurements.

Chapters 3 and 4 detailed the analysis of the MGS data sample, derived from the Sloan Digital Sky Survey. The BAO and RSD analysis of this sample is complete, and there is currently little to be gained from further efforts to improve the constraints from this sample. Indeed the MGS constraints at this redshift are unlikely to be bettered until the

DESI survey (Levi et al., 2013) looks again at this low redshift over a greater sky-area than was achieved with SDSS-II. However, the techniques used are certainly applicable to other datasets. The creation of mock catalogues is an important stage in the analysis of LSS, and allows for precise quantification of the statistical and systematic errors. Future work in this field could be the application of the mock catalogue pipeline used herein to generate mock galaxy catalogues for the eBOSS survey (Dawson et al., 2015) which has already begun taking data. Similarly, the BAO and RSD modelling and fitting techniques can be used for the analysis of this survey very easily. Looking further ahead, more work will need to be done to ensure that the RSD modelling is accurate enough on small scales to fully recover unbiased information from surveys such as Euclid. Accurate measurements of RSD in such surveys will provide unprecedented tests on the fidelity of General Relativity and the validity of modified gravity theories.

6.1.3 Improvements to the optimal covariance matrix estimation method.

Chapter 5 presented a method to scale the covariance matrix measured from a set of small volume simulations such that it matches the covariance matrix from an ensemble of large volume simulations. This opens the door for covariance matrix estimation for next generation surveys, bypasses the problem of generating the necessary simulations, and allows for an improvement in the precision of the covariance matrix. This was shown to work very well on small scales on the MGS dataset, in a regime where theoretical estimates are incredibly difficult to obtain and very inaccurate. At the moment however, this remains a proof of concept and showing that applying this technique to a Euclid like survey enables easier, more precise estimates of the covariance matrix than would otherwise be possible remains future work.

At the end of this chapter it was also shown that the scaled covariance matrix does not agree with the covariance matrix measured from a set of large simulations on large scales. This is due to the window function. Large volume simulations that cover the full survey volume can simply be masked before the covariance matrix is calculated. The resultant measurements include the effects of the convolution with the survey window. This cannot be done with the small volume mocks and so a method of including this method analytically is required, much like a model power spectrum must be convolved with a survey window function to reproduce the measured power spectrum. Future work in this area would be to use the formalism developed in Chapter 5 to investigate the effect of the window function convolution on the covariance matrix mathmatically, and to find a simple method to convolve the measured covariance matrix in a similar way to that used on the power spectrum itself.

Combined with the volume scaling technique demonstrated in this work, this would

allow one to recover the correct covariance matrix on all scales, from a set of simulations much smaller than the survey volume. The completion of this work would be incredibly useful for future LSS measurements and hence to measurements of the late time evolution and cosmology of the universe.

Bibliography

Abazajian K. N. et al., 2009, ApJS, 182, 543

Adler, R. J. 1981, The Geometry of Random Fields, Chichester: Wiley, 1981,

Ahn, C. P., Alexandroff, R., Allende Prieto, C., et al. 2012, ApJS, 203, 21

Ahn, C. P., Alexandroff, R., Allende Prieto, C., et al. 2014, ApJS, 211, 17

Albrecht, A., & Steinhardt, P. J. 1982, Physical Review Letters, 48, 1220

Alcock C., Paczynski B., 1979, Nature, 281, 358.

Alpher, R. A., Bethe, H., & Gamow, G. 1948, Physical Review, 73, 803

Anderson, L., Aubourg, E., Bailey, S., et al. 2014, MNRAS, 439, 83

Anderson, L., Aubourg, É., Bailey, S., et al. 2014, MNRAS, 441, 24

Angulo, R. E., Baugh, C. M., Frenk, C. S., & Lacey, C. G. 2008, MNRAS, 383, 755

Angulo, R. E., Baugh, C. M., Frenk, C. S., & Lacey, C. G. 2014, MNRAS, 442, 3256

Aubourg, É., Bailey, S., Bautista, J. E., et al. 2014, arXiv:1411.1074

Bardeen, J. M., Bond, J. R., Kaiser, N., & Szalay, A. S. 1986, ApJ, 304, 15

Barnes, J., & Hut, P. 1986, Nature, 324, 446

Bennett, C. L., Kogut, A., Hinshaw, G., et al. 1994, ApJ, 436, 423

Bennett, C. L., Halpern, M., Hinshaw, G., et al. 2003, ApJS, 148, 1

Bennett, C. L., Larson, D., Weiland, J. L., et al. 2013, ApJS, 208, 20

Berlind, A. A., & Weinberg, D. H. 2002, ApJ, 575, 587

Betoule, M., Kessler, R., Guy, J., et al. 2014, A&A, 568, A22

Beutler F., et al., 2011, MNRAS, 416, 3017

Beutler F., et al., 2012, MNRAS, 423, 3430

Beutler, F., Saito, S., Seo, H.-J., et al. 2013, arXiv:1312.4611

Blake, C., Brough, S., Colless, M., et al. 2011, MNRAS, 415, 2876

Blake, C., Glazebrook, K., Davis, T. M., et al. 2011, MNRAS, 418, 1725

Blanton, M. R., et al. 2003, AJ, 125, 2348

Blanton, M. R., et al. 2005, AJ, 129, 2562

Bouchet, F. R., Colombi, S., Hivon, E., & Juszkiewicz, R. 1995, A&A, 296, 575

Bryan, G. L., & Norman, M. L. 1998, ApJ, 495, 80

Burden, A., Percival, W. J., Manera, M., et al. 2014, MNRAS, 445, 3152

Carlson, J., Reid, B., & White, M. 2013, MNRAS, 429, 1674

Chandrasekhar, S. 1931, ApJ, 74, 81

Chuang, C.-H., Prada, F., Beutler, F., et al. 2013, arXiv:1312.4889

Chuang, C.-H., Kitaura, F.-S., Prada, F., Zhao, C., & Yepes, G. 2015, MNRAS, 446, 2621

Clarkson, C., & Maartens, R. 2010, Classical and Quantum Gravity, 27, 124008

Clowes, R. G., Harris, K. A., Raghunathan, S., et al. 2013, MNRAS, 429, 2910

Coc, A., Vangioni-Flam, E., Descouvemont, P., Adahchour, A., & Angulo, C. 2004, ApJ, 600, 544

Cole, S., & Kaiser, N. 1989, MNRAS, 237, 1127

Cole, S. 1997, MNRAS, 286, 38

Cole, S., Percival, W. J., Peacock, J. A., et al. 2005, MNRAS, 362, 505

Coles, P., & Barrow, J. D. 1987, MNRAS, 228, 407

Coles, P., & Jones, B. 1991, MNRAS, 248, 1

Colless, M., Dalton, G., Maddox, S., et al. 2001, MNRAS, 328, 1039

Colless, M., Peterson, B. A., Jackson, C., et al. 2003, arXiv:astro-ph/0306581

Conley, A., Guy, J., Sullivan, M., et al. 2011, ApJS, 192, 1

Couchman, H. M. P. 1987, MNRAS, 225, 777

Crittenden, R. G., & Turok, N. 1996, Physical Review Letters, 76, 575

Crocce, M., & Scoccimarro, R. 2006, Phys. Rev. D, 73, 063519

Cruz, M., Cayón, L., Martínez-González, E., Vielva, P., & Jin, J. 2007, ApJ, 655, 11

Dalal, N., Doré, O., Huterer, D., & Shirokov, A. 2008, Phys. Rev. D, 77, 123514

Das, S., Marriage, T. A., Ade, P. A. R., et al. 2011, ApJ, 729, 62

Das, S., Louis, T., Nolta, M. R., et al. 2014, JCAP, 4, 14

Davis, M., Efstathiou, G., Frenk, C. S., & White, S. D. M. 1985, ApJ, 292, 371

Dawson, K. S., Schlegel, D. J., Ahn, C. P., et al. 2013, AJ, 145, 10

Dawson, K. S., Kneib, J.-P., Percival, W. J., et al. 2015, arXiv:1508.04473

The Dark Energy Survey Collaboration 2005, arXiv:astro-ph/0510346

de la Torre, S., Guzzo, L., Peacock, J. A., et al. 2013, A&A, 557, A54

de Putter, R., Wagner, C., Mena, O., Verde, L., & Percival, W. J. 2012, JCAP, 4, 019

Delubac, T., Bautista, J. E., Busca, N. G., et al. 2014, arXiv:1404.1801

Dodelson, S. 2003, Modern cosmology / Scott Dodelson. Amsterdam (Netherlands): Academic Press. ISBN 0-12-219141-2, 2003, XIII + 440 p.,

Dodelson, S., & Schneider, M. D. 2013, Phys. Rev. D, 88, 063537

Doroshkevich, A. G., Zel'dovich, Y. B., & Syunyaev, R. A. 1978, Soviet Astronomy, 22, 523

Drinkwater, M. J., Jurek, R. J., Blake, C., et al. 2010, MNRAS, 401, 1429

Efstathiou, G. 2014, MNRAS, 440, 1138

Einasto, J., Klypin, A. A., Saar, E., & Shandarin, S. F. 1984, MNRAS, 206, 529

Einasto, J., Einasto, M., Gottlöber, S., et al. 1997, Nature, 385, 139

Einhorn, M. B., & Sato, K. 1981, Nuclear Physics B, 180, 385

Einstein, A. 1916, Annalen der Physik, 354, 769

Eisenstein, D. J., & Hu, W. 1998, ApJ, 496, 605

Eisenstein D.J., et al., 2001, AJ, 122, 2267

Eisenstein, D. J., Zehavi, I., Hogg, D. W., et al. 2005, ApJ, 633, 560

Eisenstein, D. J., Seo, H.-J., & White, M. 2007, ApJ, 664, 660

Eisenstein, D. J., Seo, H.-J., Sirko, E., & Spergel, D. N. 2007, ApJ, 664, 675

Eisenstein, D. J., Weinberg, D. H., Agol, E., et al. 2011, AJ, 142, 72

Falck, B. L., Neyrinck, M. C., & Szalay, A. S. 2012, ApJ, 754, 126

Feldman, H. A., Kaiser, N., & Peacock, J. A. 1994, ApJ, 426, 23

Fisher, K. B. 1995, ApJ, 448, 494

Fixsen, D. J., Hinshaw, G., Bennett, C. L., & Mather, J. C. 1997, ApJ, 486, 623

Font-Ribera, A., Kirkby, D., Busca, N., et al. 2014, JCAP, 5, 27

Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. 2013, PASP., 125, 306

Fosalba, P., Crocce, M., Gaztanaga, E., & Castander, F. J. 2013, arXiv:1312.1707

Freedman, W. L., Madore, B. F., Scowcroft, V., et al. 2012, ApJ, 758, 24

Frenk, C. S., White, S. D. M., Davis, M., & Efstathiou, G. 1988, ApJ, 327, 507

Friedmann, A. 1922, Zeitschrift fur Physik, 10, 377

Friedmann, A. 1924, Zeitschrift fur Physik, 21, 326

Fry, J. N. 1986, ApJ, 306, 358

Fukugita, M., Ichikawa, T., Gunn, J. E., Doi, M., Shimasaku, K., Schneider, D. P., 1996, AJ, 111, 1748

Gaztañaga, E., Cabré, A., & Hui, L. 2009, MNRAS, 399, 1663

Giocoli, C., Tormen, G., Sheth, R. K., & van den Bosch, F. C. 2010, MNRAS, 404, 502

Gunn, J. E., et al., 1998, AJ, 116, 3040

Gunn, J. E., et al. 2006, AJ, 131, 2332

Guth, A. H. 1981, Phys. Rev. D, 23, 347

Guzzo, L., Pierleoni, M., Meneux, B., et al. 2008, Nature, 451, 541

Hamilton, A. J. S. 1992, ApJL, 385, L5

Hamilton, A. J. S., Rimes, C. D., & Scoccimarro, R. 2006, MNRAS, 371, 1188

Hartlap, J., Simon, P., & Schneider, P. 2007, A&A, 464, 399

Heath, D. J. 1977, MNRAS, 179, 351

Henry, J. P., Evrard, A. E., Hoekstra, H., Babul, A., & Mahdavi, A. 2009, ApJ, 691, 1307

Heß, S., Kitaura, F.-S., & Gottlöber, S. 2013, MNRAS, 435, 2065

Heymans, C., Van Waerbeke, L., Miller, L., et al. 2012, MNRAS, 427, 146

Hikage, C. 2014, MNRAS, 441, L21

Hinshaw, G., Larson, D., Komatsu, E., et al. 2013, ApJS, 208, 19

Hockney, R. W., Eastwood, J. W., *Computer Simulation Using Particles*, 1988, Adam Hilger

Horváth, I., Hakkila, J., & Bagoly, Z. 2014, A&A, 561, L12

Howlett, C., Lewis, A., Hall, A., & Challinor, A. 2012, J. Cosmo. Astroparticle Phys., 4, 027

Howlett, C., Ross, A. J., Samushia, L., Percival, W. J., & Manera, M. 2015, MNRAS, 449, 848

Howlett, C., Manera, M., & Percival, W. J. 2015, Astronomy and Computing, 12, 109

Hu, W., & Sugiyama, N. 1995, ApJ, 444, 489

Hu, W., & White, M. 1996, ApJ, 471, 30

Hu, W., & Kravtsov, A. V. 2003, ApJ, 584, 702

Hubble, E. 1929, Proceedings of the National Academy of Science, 15, 168

Huchra, J. P., & Geller, M. J. 1982, ApJ, 257, 423

Hütsi, G. 2006, A&A, 449, 891

Ivezic, Z., Tyson, J. A., Abel, B., et al. 2008, arXiv:0805.2366

Jensen, L. G., & Szalay, A. S. 1986, ApJL, 305, L5

Jones, D. H., Saunders, W., Colless, M., et al. 2004, MNRAS, 355, 747

Jones, D. H., Read, M. A., Saunders, W., et al. 2009, MNRAS, 399, 683

Kaiser, N. 1984, ApJL, 284, L9

Kaiser N., 1987, MNRAS, 227, 1

Kaufman, C., Schervish, M., Nychka, D. 2008, J. American. Stat. Assoc., 103, 484.

Kayo, I., Takada, M., & Jain, B. 2013, MNRAS, 429, 344

Kazanas, D. 1980, ApJL, 241, L59

Kazin, E. A., Blanton, M. R., Scoccimarro, R., et al. 2010, ApJ, 710, 1444

Kazin, E. A., Koda, J., Blake, C., et al. 2014, MNRAS, 441, 3524

Keisler, R., Reichardt, C. L., Aird, K. A., et al. 2011, ApJ, 743, 28

Kitaura, F.-S., & Heß, S. 2013, MNRAS, 435, L78

Kitaura, F.-S., Yepes, G., & Prada, F. 2014, MNRAS, 439, L21

Klypin, A., & Holtzman, J. 1997, arXiv:astro-ph/9712217

Klypin, A. A., Trujillo-Gomez, S., & Primack, J. 2011, ApJ, 740, 102

Knebe, A., Knollmann, S. R., Muldrew, S. I., et al. 2011, MNRAS, 415, 2293

Kulkarni, G. V., Nichol, R. C., Sheth, R. K., et al. 2007, MNRAS, 378, 1196

Lacey, C., & Cole, S. 1994, MNRAS, 271, 676

Landy S. D., Szalay A. S., 1993, ApJ, 412, 64

Laureijs, R., Amiaux, J., Arduini, S., et al. 2011, arXiv:1110.3193

Lemaître, G. 1927, Annales de la Société Scientifique de Bruxelles, 47, 49

Levi, M., Bebek, C., Beers, T., et al. 2013, arXiv:1308.0847

Lewis, A., Challinor, A., & Lasenby, A. 2000, ApJ, 538, 473

A., Lewis, S., Bridle 2002, PRD, 66, 103511

A., Lewis 2013, PRD, 87, 103529

Lesgourgues, J. 2011, arXiv:1104.2932

Li, Y., Hu, W., & Takada, M. 2014, Phys. Rev. D, 89, 083519

Li, Y., Hu, W., & Takada, M. 2014, Phys. Rev. D, 90, 103530

Liddle, A. R., & Lyth, D. H. 2000, Cosmological Inflation and Large-Scale Structure, by Andrew R. Liddle and David H. Lyth, pp. 414. ISBN 052166022X. Cambridge, UK: Cambridge University Press, April 2000.,

Liddle, A. 2003, An Introduction to Modern Cosmology, Second Edition, by Andrew Liddle, pp. 188. ISBN 0-470-84834-0. Wiley-VCH , May 2003.,

Linde, A. D. 1982, Physics Letters B, 108, 389

Linder, E. V., & Cahn, R. N. 2007, Astroparticle Phys., 28, 481

Lumsden, S. L., Heavens, A. F., & Peacock, J. A. 1989, MNRAS, 238, 293

Lyth, D. H., & Liddle, A. R. 2009, The Primordial Density Perturbation, by David H. Lyth , Andrew R. Liddle, Cambridge, UK: Cambridge University Press, 2009,

Maartens, R., Ellis, G. F. R., & Stoeger, W. R. 1995, Phys. Rev. D, 51, 1525

Maartens, R., Abdalla, F. B., Jarvis, M., Santos, M. G., & SKA Cosmology SWG, f. t. 2015, arXiv:1501.04076

Manera, M., Scoccimarro, R., Percival, W. J., et al. 2013, MNRAS, 428, 1036

Manera, M., Samushia, L., Tojeiro, R., et al. 2015, MNRAS, 447, 437

Mantz, A., Allen, S. W., Rapetti, D., & Ebeling, H. 2010, MNRAS, 406, 1759

Mather, J. C., Cheng, E. S., Cottingham, D. A., et al. 1994, ApJ, 420, 439

Matsubara, T. 1995, Progress of Theoretical Physics, 94, 1151

Matsubara, T. 2008, Phys. Rev. D, 78, 083519

Mehta, K. T., Seo, H.-J., Eckel, J., et al. 2011, ApJ, 734, 94

Meiksin, A., & White, M. 1999, MNRAS, 308, 1179

Merson, A. I., Baugh, C. M., Helly, J. C., et al. 2013, MNRAS, 429, 556

Meszaros, P. 1974, A&A, 37, 225

Miller, C. J., Nichol, R. C., & Batuski, D. J. 2001, ApJ, 555, 68

Monaco, P., Theuns, T., Taffoni, G., et al. 2002, ApJ, 564, 8

Monaco, P., Sefusatti, E., Borgani, S., et al. 2013, MNRAS, 433, 2389

More, S., Kravtsov, A. V., Dalal, N., & Gottlöber, S. 2011, ApJS, 195, 4

Moutarde, F., Alimi, J.-M., Bouchet, F. R., Pellat, R., & Ramani, A. 1991, ApJ, 382, 377

Mukhanov, V. 2005, Physical Foundations of Cosmology, by Viatcheslav Mukhanov, pp. 442. Cambridge University Press, November 2005. ISBN-10: 0521563984. ISBN-13: 9780521563987. LCCN: QB981 .M89 2005,

Navarro J.F., Frenk C.S., White S.D.M. 1996, ApJ, 462, 563

Nusser, A., & Davis, M. 1994, ApJL, 421, L1

Oka, A., Saito, S., Nishimichi, T., Taruya, A., & Yamamoto, K. 2014, MNRAS, 439, 2515

Okumura, T., Matsubara, T., Eisenstein, D. J., et al. 2008, ApJ, 676, 889

Pacheco, P. S., *Parallel Programming with MPI*, 1997, Morgan Kaufmann.

Padmanabhan, N., et al. 2008, ApJ, 674, 1217

Padmanabhan, N., & White, M. 2009, Phys. Rev. D, 80, 063508

Padmanabhan, N., Xu, X., Eisenstein, D. J., et al. 2012, MNRAS, 427, 2132

Paz, D., Lares, M., Ceccarelli, L., Padilla, N., & Lambas, D. G. 2013, MNRAS, 436, 3480

Paz, D. J., & Sanchez, A. G. 2015, arXiv:1508.03162

Peacock, J. A., & Heavens, A. F. 1985, MNRAS, 217, 805

Peacock, J. A. 1999, Cosmological Physics, by John A. Peacock, pp. 704. ISBN 052141072X. Cambridge, UK: Cambridge University Press, January 1999.,

Pearson, D. W., & Samushia, L. 2015, arXiv:1509.00064

Peebles, P. J. E., & Yu, J. T. 1970, ApJ, 162, 815

Peebles, P. J. E. 1980, Research supported by the National Science Foundation. Princeton, N.J., Princeton University Press, 1980. 435 p.,

Peebles, P. J. E., Melott, A. L., Holmes, M. R., & Jiang, L. R. 1989, ApJ, 345, 108

Penzias, A. A., & Wilson, R. W. 1965, ApJ, 142, 419

Percival, W. J., Baugh, C. M., Bland-Hawthorn, J., et al. 2001, MNRAS, 327, 1297

Percival, W. J., Burkey, D., Heavens, A., et al. 2004, MNRAS, 353, 1201

Percival, W. J., Cole, S., Eisenstein, D. J., et al. 2007, MNRAS, 381, 1053

Percival, W. J., & White, M. 2009, MNRAS, 393, 297

Percival, W. J., Reid, B. A., Eisenstein, D. J., et al. 2010, MNRAS, 401, 2148

Percival, W. J., Ross, A. J., Sánchez, A. G., et al. 2014, MNRAS, 439, 2531

Perlmutter S. et al., 1999, ApJ, 517, 565

Phillips, M. M. 1993, ApJL, 413, L105

Planck Collaboration - I, Ade, P. A. R., Aghanim, N., et al. 2014, A& A, 571, A1

Planck Collaboration - XVI, Ade, P. A. R., Aghanim, N., et al. 2014, A& A, 571, A16

Planck Collaboration - XXIII, Ade, P. A. R., Aghanim, N., et al. 2014, A&A, 571, A23

Planck Collaboration, Ade, P. A. R., Aghanim, N., et al. 2015, arXiv:1502.01589

Planck Collaboration, Ade, P. A. R., Aghanim, N., et al. 2015, arXiv:1502.01592

Pope, A. C., & Szapudi, I. 2008, MNRAS, 389, 766

Prada, F., Klypin, A. A., Cuesta, A. J., Betancort-Rijo, J. E., & Primack, J. 2012, MNRAS, 423, 3018

Pskovskii, I. P. 1977, Soviet Astronomy, 21, 675

Quinn, T., Katz, N., Stadel, J., & Lake, G. 1997, arXiv:astro-ph/9710043

Rees, M. J., & Sciama, D. W. 1968, Nature, 217, 511

Reid, B. A., Percival, W. J., Eisenstein, D. J., et al. 2010, MNRAS, 404, 60

Reid, B. A., & White, M. 2011, MNRAS, 417, 1913

Reid, B. A., Samushia, L., White, M., et al. 2012, MNRAS, 426, 2719

Riess A. G. et al., 1998, AJ, 116, 1009

Riess, A. G., Macri, L., Casertano, S., et al. 2011, ApJ, 730, 119

Rimes, C. D., & Hamilton, A. J. S. 2006, MNRAS, 371, 1205

Robertson, H. P. 1935, ApJ, 82, 284

Ross, A. J., Percival, W. J., Sánchez, A. G., et al. 2012, MNRAS, 424, 564

Ross, A. J., Percival, W. J., Carnero, A., et al. 2013, MNRAS, 428, 1116

Ross, A. J., Samushia, L., Burden, A., et al. 2014, MNRAS, 437, 1109

Ross, A. J., Samushia, L., Howlett, C., et al. 2015, MNRAS, 449, 835

Rozo, E., Wechsler, R. H., Rykoff, E. S., et al. 2010, ApJ, 708, 645

Sachs, R. K., & Wolfe, A. M. 1967, ApJ, 147, 73

Samushia, L., Percival, W. J., & Raccanelli, A. 2012, MNRAS, 420, 2102

Samushia, L., Reid, B. A., White, M., et al. 2014, MNRAS, 439, 3504

Sánchez, A. G., Montesano, F., Kazin, E. A., et al. 2014, MNRAS, 440, 2692

Sato, M., Hamana, T., Takahashi, R., et al. 2009, ApJ, 701, 945

Schäfer, J., & Strimmer, K. 2005, Stat. App. Genet. Mol. Biol., 4, 32

Scoccimarro, R. 1998, MNRAS, 299, 1097

Scoccimarro, R., Zaldarriaga, M., & Hui, L. 1999, ApJ, 527, 1

Scoccimarro, R., & Sheth, R. K. 2002, MNRAS, 329, 629

Scoccimarro, R. 2004, Phys. Rev. D., 70, 083007

Scoccimarro, R., Hui, L., Manera, M., & Chan, K. C. 2012, Phys. Rev. D , 85, 083002

Scrimgeour, M. I., Davis, T., Blake, C., et al. 2012, MNRAS, 425, 116

Sefusatti, E., Crocce, M., Pueblas, S., & Scoccimarro, R. 2006, Phys. Rev. D, 74, 023522

Seo H.-J., Eisenstein D.J., 2003, ApJ, 598, 720

Seo, H.-J., Ho, S., White, M., et al. 2012, ApJ, 761, 13

Seljak, U., & Zaldarriaga, M. 1996, ApJ, 469, 437

Silk, J. 1968, ApJ, 151, 459

Sirko, E. 2005, ApJ, 634, 728

Slipher, V. M. 1912, Lowell Observatory Bulletin, 2, 26

Smith, R. E., & Marian, L. 2015, arXiv:1503.06830

Smoot, G. F., Bennett, C. L., Kogut, A., et al. 1992, ApJL, 396, L1

Spergel, D. N., Verde, L., Peiris, H. V., et al. 2003, ApJS, 148, 175

Splinter, R. J., Melott, A. L., Shandarin, S. F., & Suto, Y. 1998, ApJ, 497, 38

Springel, V. 2005, MNRAS, 364, 1105

Story, K. T., Reichardt, C. L., Hou, Z., et al. 2013, ApJ, 779, 86

Starobinskiĭ, A. A. 1979, Soviet Journal of Experimental and Theoretical Physics Letters, 30, 682

Strauss, M.A., et al. 2002, AJ, 124, 1810

Swanson, M. E. C., Tegmark, M., Hamilton, A. J. S., & Hill, J. C. 2008, MNRAS, 387, 1391

Takada, M., & Bridle, S. 2007, New Journal of Physics, 9, 446

Takada, M., & Jain, B. 2009, MNRAS, 395, 2065

Takada, M., & Hu, W. 2013, Phys. Rev. D, 87, 123504

Takahashi, R., Yoshida, N., Takada, M., et al. 2009, ApJ, 700, 479

Takahashi, R., Yoshida, N., Takada, M., et al. 2011, ApJ, 726, 7

Takahashi, R., Soma, S., Takada, M., & Kayo, I. 2014, MNRAS, 444, 3473

Tammann, G. A., & Reindl, B. 2013, A&A, 549, A136

Tassev, S., Zaldarriaga, M., & Eisenstein, D. J. 2013, J. Cosmo. Astroparticle Phys., 6, 36

Tassev, S., Eisenstein, D. J., Wandelt, B. D., & Zaldarriaga, M. 2015, arXiv:1502.07751

Taylor, A., Joachimi, B., & Kitching, T. 2013, MNRAS, 432, 1928

Tegmark, M. 1997, Physical Review Letters, 79, 3806

Terukina, A., Lombriser, L., Yamamoto, K., et al. 2014, JCAP, 4, 013

Tinker, J., Kravtsov, A. V., Klypin, A., et al. 2008, ApJ, 688, 709

Tojeiro, R., Ross, A. J., Burden, A., et al. 2014, MNRAS, 440, 2222

Tormen, G., & Bertschinger, E. 1996, ApJ, 472, 14

Vikhlinin, A., Kravtsov, A. V., Burenin, R. A., et al. 2009, ApJ, 692, 1060

Walker, A. G. 1935, MNRAS, 95, 263

Wang, L., Reid, B., & White, M. 2014, MNRAS, 437, 588

Waterhouse, T. P. 2006, arXiv:astro-ph/0611816

Weinberg, D. H., Mortonson, M. J., Eisenstein, D. J., et al. 2013, Phys. Rep., 530, 87

Weinberg, S. 2008, Cosmology, by Steven Weinberg. ISBN 978-0-19-852682-7. Published by Oxford University Press, Oxford, UK, 2008.,

White, M., Tinker, J. L., & McBride, C. K. 2014, MNRAS, 437, 2594

Wilcox, H., Bacon, D., Nichol, R. C., et al. 2015, MNRAS, 452, 1171

Xu, X., Padmanabhan, N., Eisenstein, D. J., Mehta, K. T., & Cuesta, A. J. 2012, MNRAS, 427, 2146

Xu, X., Cuesta, A. J., Padmanabhan, N., Eisenstein, D. J., & McBride, C. K. 2013, MNRAS, 431, 2834

York, D. G., Adelman, J., Anderson, J. E., Jr., et al. 2000, AJ, 120, 1579

Zehavi, I., Zheng, Z., Weinberg, D. H., et al. 2011, ApJ, 736, 59

Zel'dovich, Y. B. 1970, A& A, 5, 84

Zheng, Z., Coil, A. L., & Zehavi, I. 2007, ApJ, 667, 760

FORM UPR16

Research Ethics Review Checklist

University of Portsmouth

Please include this completed form as an appendix to your thesis (see the Postgraduate Research Student Handbook for more information

Postgraduate Research Student (PGRS) Information				**Student ID:**	673693
PGRS Name:	CULLAN HOWLETT				
Department:	ICG, TECH	**First Supervisor:**	PROF. WILL PERCIVAL		
Start Date: (or progression date for Prof Doc students)		**01/10/2012**			
Study Mode and Route:	Part-time ☐ Full-time ☒	MPhil ☐ PhD ☒	MD ☐ Professional Doctorate ☐		

Title of Thesis:	Modelling and Measuring Cosmological Structure Growth
Thesis Word Count: (excluding ancillary data)	63,998

If you are unsure about any of the following, please contact the local representative on your Faculty Ethics Committee for advice. Please note that it is your responsibility to follow the University's Ethics Policy and any relevant University, academic or professional guidelines in the conduct of your study

Although the Ethics Committee may have given your study a favourable opinion, the final responsibility for the ethical conduct of this work lies with the researcher(s).

UKRIO Finished Research Checklist:
(If you would like to know more about the checklist, please see your Faculty or Departmental Ethics Committee rep or see the online version of the full checklist at: http://www.ukrio.org/what-we-do/code-of-practice-for-research/)

a) Have all of your research and findings been reported accurately, honestly and within a reasonable time frame?	YES ☒ NO ☐
b) Have all contributions to knowledge been acknowledged?	YES ☒ NO ☐
c) Have you complied with all agreements relating to intellectual property, publication and authorship?	YES ☒ NO ☐
d) Has your research data been retained in a secure and accessible form and will it remain so for the required duration?	YES ☒ NO ☐
e) Does your research comply with all legal, ethical, and contractual requirements?	YES ☒ NO ☐

Candidate Statement:

I have considered the ethical dimensions of the above named research project, and have successfully obtained the necessary ethical approval(s)

Ethical review number(s) from Faculty Ethics Committee (or from NRES/SCREC):	0B0E-E088-ED86-D891-9CA0-610E-0F5E-4D20

If you have *not* submitted your work for ethical review, and/or you have answered 'No' to one or more of questions a) to e), please explain below why this is so:

Signed *(PGRS)*: 	**Date:** 21/03/2016